Post-transcriptional Regulation of Gene Expression by Small RNAs and RNA-binding Proteins

By

Ting Han

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Cell and Developmental Biology)
in The University of Michigan
2013

Doctoral Committee:

Assistant Professor John Kim, Chair
Professor John Moran
Professor Lois Weisman
Assistant Professor Kentaro Nabeshima

ACKNOWLEDGMENTS

I could never imagine being where I am without the tremendous support and guidance from my advisor, Prof. John Kim. John's enthusiasm, vision, attention to details, and generosity have been and will continue to be a great source of aspiration for my own scientific inquiries. His dedication to my scientific development has left an indelible mark on me. I will strive to pass down my scientific heritage from him to my future trainees.

In addition to John's mentorship, I am blessed to have an outstanding thesis committee, which includes Drs. John Moran, Lois Weisman, and Kentaro Nabeshima; all with great knowledge and patience. I am grateful for their encouragement, insightful comments, and valuable career advice.

This thesis would not have been possible without the hard work of my colleagues Vishal Khivansara, Mallory Freeberg, and Arun Prasad Manoharan, who helped me enormously in all of the projects described in the thesis. Allison Billi spent countless hours editing my manuscripts and has been a great sounding board for new scientific ideas. I am also thankful to other fellow lab members, including Danny Yang, Tony Chun, Natasha Weiser, Amelia Alessi, Da Fang, and past lab members Amanda Day, Matthew Avenarius, and Dongping Wei, who made the past six years I spent in the Kim lab very enjoyable.

My sincere thanks go to many collaborators who have generously shared their expertise and help me complete the projects in this thesis. Diana Chu (San Francisco State University) spent months purifying germ cells from *C. elegans*. Without her effort, I would not have discovered 26G RNAs. James Moresco (The Scripps Research Institute) performed many mass spectrometry analyses, which constitute the cornerstone of several projects in my thesis. I would also like to thank the yeast community on the sixth floor in LSI, particularly Lois Weisman for sharing yeast GFP and TAP collections, and Meiyan Jin, Ke Wang, and Dan Klionsky for reagents and advice.

I am indebted to my parents, family members, and friends; many of them are thousands of miles away. In particular, I am beyond grateful to my mother, Hanzhu Jia, whose love, patience, understanding and encouragement have always been with me no matter how far I go.

PREFACE

This thesis comprises the research I conducted in Prof. John Kim's lab starting from June 2007. The objective was to better characterize the mechanisms by which small RNAs and RNA-binding proteins (RBPs) govern post-transcriptional regulation of gene expression.

Chapters 2, 4 and 5 were published in *PNAS* (2009;106(44):18674-9), *Science* (2010;329(5990):432-5), and *Genome Biology* (2013;14(2):R13), respectively. Chapter 3 is currently under revision at PNAS and will be submitted by April 2013.

Chapter 2 describes the identification of a novel class of endogenous, germline-generated small RNAs, 26G endo-siRNAs, which regulate gene expression during spermatogenesis and zygotic development. I initiated this project and performed all of the experiments described in the publication. Arun Prasad Manoharan, Danielle Thierry-Mieg, and Jean Thierry-Mieg contributed to the bioinformatic analyses; Tim Harkins and Pascal Bouffard performed 454 sequencing of the small RNA libraries; Colin Fitzpatrick and Diana Chu provided purified sperm and oocyte samples.

Chapter 3 describes the discovery that a highly conserved protein kinase, casein kinase 2, regulates target binding and repression by the miRNA-induced silencing complex (miRISC). Vishal Khivansara and I contributed equally to this work. I initiated the project, made the original observations, and established most

of the phenotypic assays. Vishal repeated many of these phenotypic assays and established all of the biochemical assays described. James Moresco, Patricia Tu, and John Yates contributed mass spec analyses of miRISC composition and phoshorylation. Mallory Freeberg performed bioinformatic analysis of miRNA expression levels.

Chapter 4 describes a collaborative effort to capture and sequence the 3' UTRs of the majority of *C. elegans* mRNAs, providing accurate 3' UTR annotations for 85% of genes and revealing the dynamic nature of alternative 3' UTR expression across the major developmental stages of *C. elegans*. This work was conceived and supervised by John Kim in collaboration with Fabio Piano and Kris Gunsalus as part of the modEncode (model organism ENCyclopedia Of DNA Elements) Consortium. Marco Mangone from Fabio Piano's group performed gene-specific 3' RACE (3' Rapid Amplification of cDNA Ends) and Sanger sequencing. I developed the PolyA capture protocol and prepared all of the 454 libraries. Pascal Bouffard and Tim Harkins performed 454 sequencing of the libraries. Yutaka Suzuki, Sumiyo Sugano, and Yuji Kohara contributed the full-length cDNA sequences. Arun Prasad Manoharan, Kris Gunsalus, Danielle Thierry-Mieg, Jean Thierry-Mieg, and Sebastian Mackowiak performed the bioinformatic analyses.

Chapter 5 investigates the global changes of RBP-RNA interactions as a response to nutrient limitation in the budding yeast *Saccharomyces cerevisiae*. This work represents the first comparative study of the *in vivo* RNA-protein interactome, and provides a comprehensive map of RBP occupancy on mRNAs.

Mallory Freeberg and I contributed equally to this work. I designed the study, developed the gPAR-CLIP methodology and performed all of the experiments. Mallory performed the bioinformatic analyses. Andy Kong contributed to the RNA secondary structure analyses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Eukaryotic mRNAs reside in ribonucleoprotein (RNP) particles that display a wide-ranging array of biochemical, subcellular, and functional properties. For example, the translational efficiencies of different mRNAs in budding yeast *Sacharomyces cerevisiae* can vary by over 100 fold (1). Similarly, the half-lives of yeast mRNAs range from ~3 min to over 90 min (2). Moreover, large-scale RNA Fluorescence In Site Hybridization (FISH) studies in fly embryos revealed that ~70% of mRNAs examined display diverse patterns of subcellular localization (3). These “personalized” properties of mRNAs can be largely explained by two types of mechanisms. One is the intrinsic difference in mRNA sequences and structures. For example, differences in 5’ UTR elements, such as secondary structures, upstream open reading frames (uORF), and sequence context of the AUG start codon, affect the interaction of mRNAs with translational machinery, which, in turn, affects their rates of translation (4). The other mechanism is post-transcriptional regulation mediated by small RNAs and RNA-binding proteins (RBP) (5, 6). For example in yeast, stability elements in the 3’ UTRs of mRNAs encoding mitochondria-localized proteins are recognized by an RBP Puf3, which recruits nucleases to promote mRNA decay (7).

Ongoing intense genome and transcriptome sequencing efforts have revealed thousands of small RNAs and hundreds of RNA-binding proteins (RBP) encoded in eukaryotic genomes. Yet, the function of only a small subset of these small RNAs and RBPs is well understood. The objective of this thesis was to understand the mechanisms by which small RNAs and RBPs govern post-transcriptional regulation of gene expression. The following introduction broadly summarizes the scope of the gene regulatory mechanisms orchestrated by small RNAs and RBPs. Detailed aspects of each area are elaborated upon in the introduction and discussion sections of subsequent chapters of the thesis.

1.1 Classes of eukaryotic small RNAs

Diverse classes of small non-coding RNAs have emerged as essential regulators of gene expression in all eukaryotes (8, 9). Small RNAs are defined by their mechanism of biogenesis, the effector proteins with which they associate, and their biological functions. microRNAs (miRNAs) are processed from double-stranded hairpin precursors by the RNase III-like enzyme Dicer to the ~22nt mature form. They associate with Argonaute (Ago) proteins in the RNA-induced silencing complex (RISC) to mediate translational repression or degradation of their target mRNAs (5, 8). Piwi-interacting RNAs (piRNAs) are generated by a Dicer-independent mechanism. They associate with Piwi proteins in RISC and silence foreign genetic elements (10). Endogenous small interfering RNAs (endo-siRNAs) represent an emerging class of small RNAs (11). In many eukaryotes, endo-siRNAs originate from endogenous genomic loci via diverse biogenesis

pathways, and regulate endogenous gene expression and epigenetic states of their loci (12-14). In the following sections, I will discuss the biogenesis and function of these three major classes of small RNAs, with a primary focus on *C. elegans* small RNA biology.

1.1.1 miRNAs

Since the discovery of the first two miRNAs, *lin-4 (15)* and *let-7 (16)*, in *C. elegans*, the members of the miRNA superfamily have emerged as conserved ubiquitous regulators of gene expression, critical for animal development, cell differentiation, apoptosis, and metabolism (5). The mechanisms of miRNA biogenesis are well understood (5, 8, 17). They are transcribed as long primary transcripts and processed by the RNase III enzyme, Drosha, into ~70 nt precursor hairpins in the nucleus. Precursors are then exported to the cytosol and processed into ~22nt duplexes by Dicer (8, 17). One strand of the duplex (the mature miRNA) is selectively loaded into an Argonaute family protein that forms the core of miRISC. Through partial base pairing with target sites predominantly in the 3' UTRs of mRNAs, miRNAs direct miRISC to target mRNAs for translational inhibition and/or target mRNA degradation (18).

A unifying mechanism of target silencing remains to be resolved with multiple studies suggesting target repression at various steps in translation, deadenylation, and mRNA decay (19-23). Recent studies that monitor the temporal effects of miRNA-mediated regulation show that all these mechanisms

contribute to silencing: miRNAs initially repress translation by reducing the rate of initiation, followed by deadenylation and mRNA decay (24, 25).

1.1.2 piRNAs

Piwi proteins bind a class of longer small RNAs, piRNAs (Piwi-interacting RNAs) that are distinct from the canonical ~22 nucleotide (nt) small RNAs (26). In *D. melanogaster*, these 24-27nt piRNAs were found to be enriched in the testes and ovaries and were derived from transposons and other repetitive elements (27, 28). Similarly, in mammals, a class of 26-31nt piRNAs was identified in both male and female germlines (29-31).

The biogenesis of piRNAs is not fully understood. Two general mechanisms have been proposed: long precursor processing and the "Ping Pong" cycle (32). Some piRNAs, including mouse pachytene piRNAs and primary piRNAs, are processed from a long precursor transcript. These piRNAs have a 5' uridine bias. For example in *D. melanogaster*, the *flamenco* locus encodes a long transcript, which is processed into piRNAs to silence the *gypsy* transposon (33). Other piRNAs are produced through a combination of both precursor processing and the "Ping Pong" cycle. During the "Ping Pong" cycle, target cleavage by primary piRNAs triggers the production of secondary piRNAs from the target, whose 5' ends correspond to the cleavage sites (33, 34). The endonucleolytic "slicer" activity of Piwi proteins is required for this cleavage event to generate the 5' ends (34). The nuclease that generates the 3' ends of piRNAs has not been identified (35). In addition, both primary piRNA processing and the

"Ping Pong" cycle appear to be broadly conserved in metazoans with the exception of nematodes (36).

The piRNAs in *C. elegans* are distinct from their counterparts in flies and mammals. *C. elegans* piRNAs, referred to as 21U RNAs, are 21 nt long and start with a 5' terminal uridine (37-39). They are transcribed as single units from two broad clusters on chromosome IV and silence foreign genetic sequences as well as endogenous genes via imperfect complementary (up to 3 mismatches) (40-43). Little is known about the exact mechanisms of 21U RNA biogenesis, except that PRG-1, the *C. elegans* Piwi ortholog, is required for their expression (38, 39). Several recent studies suggest 21U RNA biogenesis is driven by an 8 base pair core motif located ~40 base pair upstream from the start of the 21U RNA sequence (37, 38, 42, 43). In addition, 21U RNAs are likely transcribed by RNA Pol II as capped precursors that initiate 2 nt upstream of the first uridine (42, 43). However, the exact identities of these putative piRNA precursors as well as the mechanisms of their processing have not been fully uncovered.

1.1.3 endo-siRNAs

C. elegans endo-siRNAs are divided into several subclasses with unique yet interconnected functions. We (Chapter 2; (44)) and others (45, 46) characterized a class of germline-generated primary endo-siRNAs with 26 nt length and 5' terminal guanosines (i.e., 26G RNAs). Analysis of genetic mutants indicates that 26G RNAs are generated by a template-dependent mechanism and require the RRF-3 RNA-dependent RNA polymerase (RdRP) and the ERI-1 exonuclease for their biogenesis (Figure 1.1). 26G RNAs fall into two non-

overlapping sub-classes: Class I 26G endo-siRNAs associate with the Argonaute ERGO-1 and are expressed exclusively during spermatogenesis, while the class II 26G endo-siRNAs associate with the Argonautes ALG-3 and ALG-4 and are generated during oocyte development and maternally deposited in fertilized embryos of the next generation. Loss of class I 26G RNAs results in severe defects in spermatogenesis and leads to sterility, whereas loss of class II 26G RNAs enhances silencing efficacy to foreign dsRNAs.

The second class of *C. elegans* endo-siRNAs, 22G RNAs, are 22 nt long and start with 5' guanosines (37). Unlike all other small RNAs which possess 5' monophosphates, 22G RNAs are 5' triphosphorylated (37). Two subclasses of 22G RNAs can be differentiated: WAGO 22G RNAs and CSR-1 22G RNAs (13, 47). WAGO 22G RNAs are secondary siRNAs, produced from transcripts that are targeted by primary siRNAs, such as 26G RNAs, 21U RNAs, and exogenous siRNAs (Figure 1.1). They require the RdRP RRF-1 and EGO-1 for their biogenesis, and associate with twelve worm-specific Argonautes (WAGO-1 to -12). WAGO 22G RNAs exert their silencing activity through both post-transcriptional and transcriptional mechanisms (13). Two of the WAGO class Argonautes NRDE-3 and HRDE-1 function in the nucleus and mediate histone H3 Lysine 9 trimethylation, resulting in transgenerational silencing (14, 41, 48-51). CSR-1 22G RNAs are produced by the RdRPs RRF-1 and EGO-1, and associate with a single Argonaute CSR-1 (47). These 22G RNAs are produced by germline-expressed transcripts and are required for proper chromosome organization. A recent study suggested that CSR-1 is required for the 3' end

processing of histone mRNAs (52). However, the exact mechanisms of CSR-1 22G RNA biogenesis and function remain incompletely understood.

In summary, *C. elegans* encodes many classes of small RNAs, with dedicated distinct mechanisms of biogenesis and function. These diverse small RNA pathways do not function in isolation. For example, the silencing of 26G RNA target mRNAs involves two rounds of RdRP activities generating both 26G RNAs and 22G RNAs (Figure 1.1). 26G RNAs act as the primary trigger and 22G RNAs act as downstream secondary siRNAs to amplify the silencing signals (45). Similarly, the piRNA pathway also collaborates with the WAGO 22G RNA pathway to provide genome surveillance (41, 48, 50, 53).

1.2 RNA-binding proteins and their RNA substrates

In addition to small RNAs, a diverse and expanding repertoire of RNA-binding proteins (RBPs) ensures faithful expression and function of substrate mRNAs (54-56).

1.2.1 Function of RNA-binding proteins

All mRNAs are organized by RBPs and other protein co-factors into higher-order ribonucleoprotein (RNP) assemblies. These RNP structures fulfill critical functions in the biogenesis, transport, inheritance, storage, and degradation of RNA (57, 58). For example, during transcription, nascent RNAs associate with a host of nuclear RBPs into hnRNP (heterogenous nuclear ribonucleoprotein) complexes to facilitate mRNA export out of nucleus (59). In

the cytoplasm, translation rates of mRNAs are modulated by a large number of RBPs, such as polyA binding protein (PABP), various eukaryotic translation initiation factors (eIFs), and sequence-specific RBPs (Figure 1.2). Many RNAs also display specific localization mediated by cellular motor proteins and adaptor RBPs (Figure 1.2). For example, *ASH1* mRNAs are transported to the daughter cell during cell division by a motor protein Myo4 and an adaptor RBP She2 in budding yeast (60). RNAs and RBPs can also reversibly aggregate into granules, such as P-bodies and stress granules, to allow RNA storage and decay in response to stimuli (61, 62). Recent discoveries suggest that aggregation of RNA-binding proteins might be a driving force for RNA granule formation (61). Kato *et al*. found that a large fraction of RNA-binding proteins harbor low complexity (LC) sequence domains that exhibit very low diversity in primary sequence. LC domains are disordered and can be induced to form amyloid-like aggregates. Upon aggregation, LC domain-containing RBPs then bring substrate RNAs and other RBPs into granules. The processes described above and many others are driven by large, complex networks of protein-RNA interactions that provide specificity in gene regulation and fidelity in RNP assembly.

1.2.2 Identification of the substrates of RNA-binding proteins

Studies of RBP-RNA interactions have historically relied on the identification of target transcripts bound by individual RBPs. *In vitro* selection of RNA sequences that bind RBPs with high affinity (systematic evolution of ligands by exponential enrichment, SELEX) can identify primary sequence recognition

elements. For example, Nova proteins, which regulate mRNA splicing in neurons, recognize the RNA consensus sequence YCAY (63), and Y box-binding protein-1, a member of the cold shock/Y box domain protein family, recognizes a CAYC RNA motif (64). Yet these and other primary sequence elements identified *in vitro* are generally short and degenerate and appear too frequently in the transcriptome to produce biologically meaningful and experimentally practical sets of targets to validate from *in silico* target predictions. Microarray profiling of transcripts that co-purify with interacting proteins (RIP-Chip) has been widely used to detect transcripts stably associated with RBPs, such as mRNAs bound by translational components HuB, eIF-4E, and PABP in P19 embryonal carcinoma stem cells (65). Similarly, RIP-Chip experiments have identified mRNAs associated with 40 yeast RBPs and uncovered a set of potential RBP recognition motifs (66, 67), some of which were validated *in vitro* using SELEX (68). Although capable of identifying mRNA targets for select RBPs, RIP-Chip is prone to artifacts including RBP-RNA dissociation and re-association after cell lysis (69), isolation of non-specific RNAs, and indirect binding through other co-purified RBPs (70). In addition, RIP-Chip cannot detect transient interactions or resolve the exact RBP binding sites on identified transcripts.

To identify transcriptome-wide footprints of RBPs *in vivo*, UV crosslinking has been coupled with immunopurification of RBPs (CLIP) (71, 72). CLIP takes advantage of the photoreactivity of RNA bases, most often pyrimidines, with interacting amino acid side chains upon 254 nm UV irradiation (73). The formation of covalent linkages allows stringent purification of RBP-RNA

complexes and subsequent identification of crosslinked RNA fragments via cDNA sequencing. Recently, a modified CLIP technique, PAR-CLIP (photoactivatable-ribonucleoside-enhanced CLIP), has been introduced in which photoactivatable-ribonucleoside analogs are incorporated into the transcriptome in live cells to enable efficient crosslinking using 365 nm UV irradiation (74). Recent studies employing CLIP in mouse brain (75) and *C. elegans* (76) and PAR-CLIP in human embryonic kidney cells (74) have successfully decoded *in vivo* microRNA-mRNA interactions by identifying RNAs bound to Argonaute, a main component of the microRNA-induced silencing complex. PAR-CLIP has also been implemented to elucidate the regulatory mechanisms of the human antigen R (HuR) protein, which stabilizes gene expression by binding to AU-rich elements (77, 78), and to identify the transcriptome-wide distribution of non-poly(A) termination factors in yeast (79). In addition to enabling efficient crosslinking, PAR-CLIP generates frequent and non-random nucleotide (nt) substitutions at crosslinking sites to reveal specific RBP-RNA contact sites with nucleotide resolution. CLIP and PAR-CLIP techniques overcome the aforementioned caveats in the RIP-Chip approach, and provide high-resolution maps of direct RBP binding sites on the entire transcriptome, thus facilitating mechanistic studies of RBP function.

1.2.3 Identification of novel RNA-binding proteins

Two recent studies introduced the use of UV crosslinking with oligo(dT) pull-down of mRNAs followed by tandem mass spectrometry to globally identify

mRNA-binding proteins in human cell lines (80, 81). In addition to identifying known RBPs, these studies identified 315 (80) and 245 (81) novel RBPs that lack canonical RNA-binding domains and functional annotation as RNA-binding proteins. Castello et al. (80) found that RBP amino acid sequences are more disordered than those of non-RBPs and identified new classes of RNA-binding domains. Baltz et al. (81) additionally captured and sequenced protein-bound mRNAs, providing a transcriptome-wide map of potential *cis*-regulatory elements. Despite these recent advances towards understanding global protein-RNA interactions, the dynamic nature of RBP-RNA associations *in vivo* and the general principles driving these associations remain unexplored.

1.3 Summary

Small RNAs and RNA-binding proteins are important post-transcriptional regulators of gene expression. The elaborate network consisting of protein-RNA, and RNA-RNA interactions ensures fine control of gene expression to confer robustness and adaptability to developmental and environmental changes.

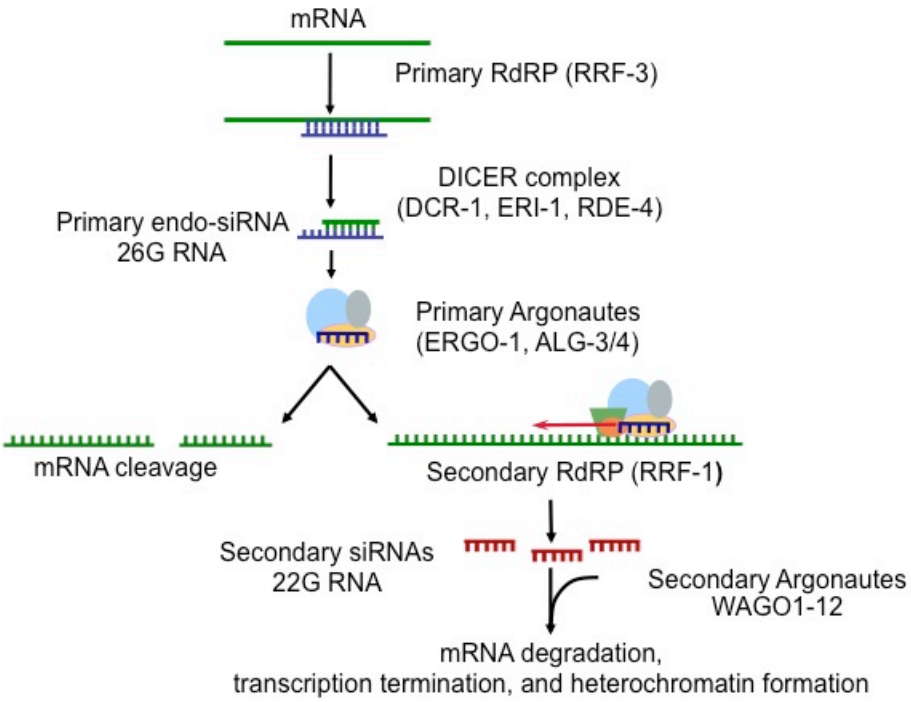


Figure 1.1 Biogenesis and functions of 26G RNAs and their downstream 22G RNAs in *C. elegans*.
26G endo-siRNAs are generated by a template-dependent mechanism and require the RRF-3 RNA-dependent RNA polymerase (RdRP) and DICER complex (containing DCR-1, ERI-1, and RDE-4) for their biogenesis. After Dicer processing, 26G RNAs associate with primary Argonautes (ERGO-1 and ALG-3/4) to cleave target mRNAs, and recruit the secondary RdRP RRF-1 to generate 22G RNAs. 22G RNAs associate with worm-specific Argonautes (WAGO-1 to -12) to elicit mRNA degradation, transcription termination, and heterochromatin formation.

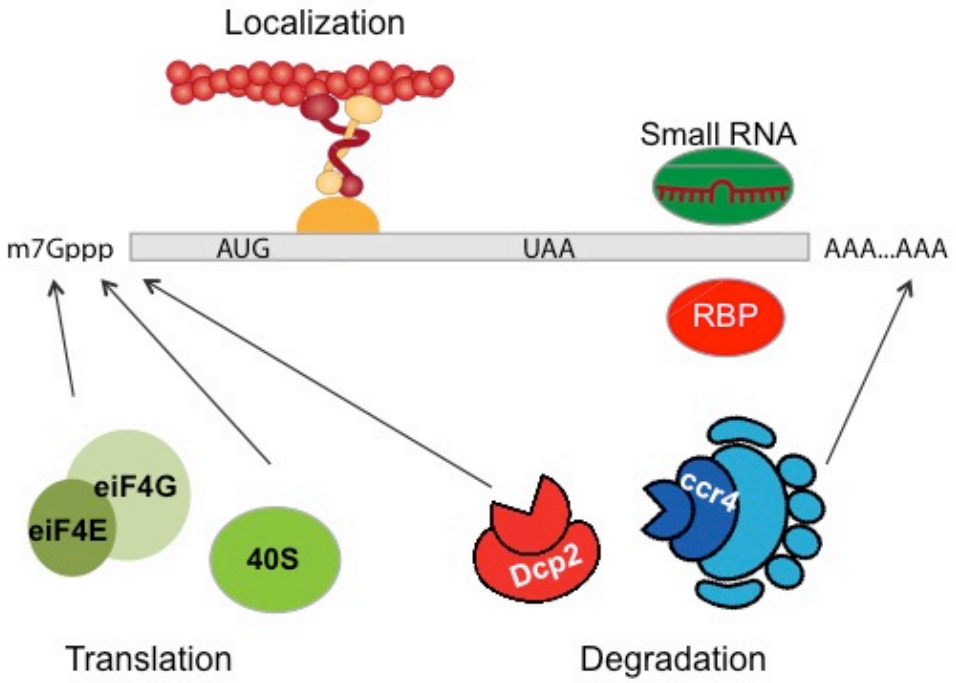


Figure 1.2 Modes of post-transcriptional regulation in eukaryotes.
Eukaryotic mRNAs interact with diverse cellular machineries, including translational initiation factors (eIFs) that decode regulatory information in 5'UTRs to determine the rate of translation, various nucleases that attack different regions of mRNAs to control their stability, motors and anchors that transport mRNAs within the cell and enrich them at various subcellular locations. Built upon these general machineries are thousands of small RNAs and hundreds of RNA-binding proteins, which recognize subsets of mRNAs for selective regulation. This figure is adapted from Roy Parker's lecture presentation on *mRNA Localization, Translation and Degradation* (http://www.ibioseminars.org/).

Chapter 2

26G endo-siRNAs regulate spermatogenic and zygotic gene expression in *C. elegans*[1]

2.1 Abstract

Endogenous small interfering RNAs (endo-siRNAs) regulate diverse gene expression programs in eukaryotes by either binding and cleaving mRNA targets or mediating heterochromatin formation; however, the mechanisms of endo-siRNA biogenesis, sorting, and target regulation remain poorly understood. Here we report the identification and function of a specific class of germline-generated endo-siRNAs in *C. elegans* that are 26nt in length and contain a guanine at the first nucleotide position (*i.e.* 26G RNAs). 26G RNAs regulate gene expression during spermatogenesis and zygotic development, and their biogenesis requires the ERI-1 exonuclease and the RRF-3 RNA-dependent RNA polymerase (RdRP). Remarkably, we identified two non-overlapping subclasses of 26G RNAs that sort into specific RNA-induced silencing complexes (RISCs) and differentially regulate distinct mRNA targets. Class I 26G RNAs target genes are expressed during spermatogenesis, whereas Class II 26G RNAs are maternally

1 Originally published in *PNAS* (2009;106(44):18674-9) with authors listed as Ting Han, Arun Prasad Manoharan, Tim T. Harkins, Pascal Bouffard, Colin Fitzpatrick, Diana S. Chu, Danielle Thierry-Mieg, Jean Thierry-Mieg, and John K. Kim.

inherited and silence gene expression during zygotic development. These findings implicate a novel class of endo-siRNAs in the global regulation of transcriptional programs required for fertility and development.

2.2 Introduction

Small RNAs bind Argonaute/Piwi proteins in the RNA-induced silencing complex (RISC) and, through base pairing, guide RISC to silence their cognate targets. While the taxonomy of small RNAs remains fluid, they can be defined in part by nucleotide length, 5' nucleotide composition, chemical modifications, genetic requirements for biogenesis, mode of silencing, and biological functions. For example, microRNAs are processed from double-stranded hairpin precursors by the RNase III-like enzyme Dicer to the ~22nt mature form containing a 5' monophosphate nucleotide. The microRNAs associate with Argonaute (Ago) proteins in RISC and mediate translational repression or degradation of their target mRNAs (5). In contrast, Piwi-interacting RNAs (piRNAs) are typically longer than microRNAs, possess a uridine in the first nucleotide, and are generated by a Dicer-independent self-amplification pathway. The piRNAs bind to Piwi proteins in RISC and silence transposons (82).

Endogenous small interfering RNAs (endo-siRNAs) represent an emerging class of small RNAs first described and characterized in *C. elegans* by Ambros *et al.* (83). These endo-siRNAs are perfectly antisense to target transcripts and require the *C. elegans* Dicer, DCR-1, the RdRP RRF-3, and the exonuclease ERI-1 for expression (84, 85). By large-scale pyrosequencing,

Ruby *et al.* determined that other endo-siRNAs target transcripts associated with spermatogenesis and transposons (37). Therefore, *C. elegans* endo-siRNAs appear to be a diverse class of small RNAs, with distinct biological functions and genetic requirements for biogenesis. The recent discovery of endo-siRNAs derived from transposable elements, natural antisense transcripts, and hairpin RNAs in *D. melanogaster* and *M. musculus* (12, 86-90) further supports their function in regulating endogenous gene expression.

Mutations affecting small RNA pathways frequently are associated with defective gametogenesis (91, 92). In *C. elegans*, mutation of *dcr-1* results in severe defects in germline development, malformed unfertilized oocytes, and sterility (92-94). Similarly, mutation of *prg-1* (*p*iwi *r*elated *g*ene) abrogates the expression of 21U RNAs (a piwi-interacting class of small RNAs) and results in impaired germline proliferation and sterility at elevated temperatures (39, 95, 96). Small RNAs also can serve as heritable parental silencing factors to regulate filial gene expression; in *D. melanogaster*, misregulation of maternally inherited piRNAs results in activation of transposons and hybrid dysgenesis (97). These observations underscore the essential functions of small RNAs in germline development and cross-generational epigenetic regulation.

In this study, we identified two classes of germline-generated endo-siRNAs, the class I sperm 26G RNAs and the class II oocyte/embryo 26G RNAs, that regulate the expression of distinct sets of genes during spermatogenesis and zygotic development, respectively. Our findings indicate that the 26G endo-siRNAs not only exert a profound influence over male gametogenesis, but also

are maternally inherited and act as epigenetic agents to control gene expression during zygotic development in the progeny.

2.3 Results

2.3.1 Deep sequencing revealed germline-enriched, *eri-1*-dependent 26G endo-siRNAs

Small RNAs expressed in purified male sperm, hermaphrodite oocytes, and embryos were size selected (18-32nt) and sequenced by high-throughput deep sequencing (Roche/454 and Illumina/Solexa). After excluding sequences corresponding to microRNAs, 21U RNAs, and putative degradation products derived from abundant noncoding RNAs (e.g. rRNAs) (Figure 2.5; Supplemental Methods), we identified 2.45 million putative endo-siRNA reads (14.8% of the total sequences). These endo-siRNAs display a bimodal length distribution with one peak clustered at ~21nt and the second at 26nt (Figure 2.1A). Notably, while the ~21nt endo-siRNAs have a first nucleotide bias for uridine observed for piRNAs in other organisms (29, 33, 98), the 26nt endo-siRNAs preferentially start with a guanine nucleotide (Figure 2.1B). Therefore, we refer to them as 26G RNAs.

Although 26G RNAs previously have been identified by deep sequencing of small RNAs isolated from mixed-stage N2 worms (37), little is known about their biogenesis or role in gene regulation. Mapping to the genome reveals that most 26G RNAs target protein coding genes (*i.e.* exons, introns, and UTRs) (77%) and exhibit a strong antisense bias (73% antisense vs. 4% sense) (Figure

2.1C; Supplemental Computational Methods). In addition, the majority of 26G RNAs are derived from exons or introns of coding transcripts target exons (97.2%) or span exon-exon junctions (0.7%), suggesting that mature mRNAs are the main targets of 26G RNAs. (Figure 2.6C).

We next used deep sequencing to compare the endo-siRNA profiles of *N2*, *glp-4(bn2)* (99), and *eri-1(mg366)* (100) whole animals. The *glp-4(bn2)* mutant fails to proliferate its germline at non-permissive temperature (25°C) and therefore lacks germline-derived small RNAs. The *glp-4* mutant exhibits a ~50% decline in 21nt siRNA expression, but a complete loss of 26G RNAs, suggesting that 26G RNAs are exclusively derived from the germline (Figure 2.1D). The *eri-1(mg366)* mutant also completely lacks 26G RNAs without globally affecting 21nt endo-siRNA levels (Figure 2.1D). Interestingly, we found a small fraction of ~21nt endo-siRNAs (4.5% of total 21nt endo-siRNAs) that appear also to be *eri-1*-dependent. These small RNAs largely overlap with 26G RNAs, starting with the same 5'G nucleotide (Figure 2.1F), and appear indistinguishable in the genetic requirements for their biogenesis from 26G RNAs. Yet, they are markedly less abundant than 26G RNAs (6.6% of the total number of 26G RNAs) (Figure 2.1F). These findings suggest that, while ~21nt endo-siRNAs, as a whole, constitute a genetically diverse population of small RNAs, 26G RNAs represent a class of germline-enriched endo-siRNAs that exclusively depends on both germline development and *eri-1* for their expression.

2.3.2 Two subclasses of 26G RNAs exhibit different expression patterns

Strikingly, hierarchical clustering reveals that 98.9% of the 26G RNAs fall into two distinct classes (Figure 2.2A; Figure 2.5; Supplemental Computational Methods). Class I 26G RNAs are present in purified sperm (1,102 unique sequences; 5,960 total reads), but are not detectable in oocytes or embryos. By comparison, class II 26G RNAs are highly enriched in oocytes and embryos (2,441 unique sequences; 148,594 total reads), but are not readily detected in sperm. Both classes of 26G RNAs are present at lower levels in mixed-stage N2 and are severely depleted in glp-4(bn2) and eri-1(mg366) animals. We analyzed the expression profiles of four relatively abundant sperm 26G RNAs (26G-S1, -S4, -S5, -S6) and four oocyte/embryo 26G RNAs (26G-16, -O1, -O2, -O3) by northern blotting and/or RT-qPCR assays (Taqman, Applied Biosystems). The stem-loop structure of the Taqman primers specifically recognizes the 3' ends of the 26G RNAs for reverse transcription and, therefore, allows discrimination from the ~21nt endo-siRNAs that start with the same 5'G as the 26G RNAs. Northern blotting demonstrated that the expression of 26G RNAs is dependent on eri-1 in purified oocytes and embryos as well as in male animals (Figure 2.2B). In addition, clear temporal separation in the expression of these two classes of 26G RNAs was observed (Figure 2.2C-D). The class I sperm 26G RNAs (denoted 26G-S) are only detectable in late larval (L4) and young adult stages in N2 hermaphrodites and males (Figure 2.2C; top panel); furthermore, a finer time course revealed class I sperm 26G RNA expression occurs in a relatively narrow window, consistent with expression during spermatogenesis (Figure 2.2D).

Conversely, expression of class II oocyte/embryo 26G RNAs (denoted 26G-O) (Figure 2.2C (bottom panel); Figure 2.2D) initiates during oogenesis, peaks in embryos, and progressively declines throughout the four larval stages. Consistent with the deep sequencing data, northern blotting indicates cross-hybridization of the 26G RNA probes to a shorter ~21nt species (Figure 2.2B,C).

2.3.3 Two subclasses of 26G RNAs silence distinct sets of targets

The 26G RNAs are perfectly complementary to their predicted gene targets, suggesting that they may act as canonical siRNAs to direct the cleavage of their mRNA targets. Importantly, 26G RNAs target a different set of genes from those targeted by shorter length (20-24nt) endo-siRNAs (Figure 2.7). Because the expression patterns of the two classes of 26G RNAs are mutually exclusive, we next asked if they differentially regulate non-overlapping, discrete classes of target genes. Indeed, based on existing germline gene expression profiles (101), we found that predicted targets of class I sperm 26G RNAs are enriched 7-fold for genes expressed during spermatogenesis, whereas targets of class II oocyte/embryo 26G RNAs are depleted of all three classes of germline genes (spermatogenesis, oogenesis, and germline-intrinsic) (Figure 2.3 B). Because mutations in *eri-1* abolish the expression of both classes of 26G RNAs, we used RT-qPCR to analyze the relative expression of putative 26G RNA targets in *eri-1(mg366)* and N2 at the following five developmental time points: embryos, and 8 hrs (L1), 30 hrs (L3), 42 hrs (L4), and 70 hrs (adult) post hatching (Figure 2.3A). While transcript levels of genes not targeted by 26G

RNAs were similar in *eri-1(mg366)* and N2 animals (Figure 2.3A, bottom panel), transcripts corresponding to 11 of the 12 genes that are targeted by class I sperm 26G RNAs and all 11 genes targeted by class II oocyte/embryo 26G RNAs are significantly elevated in *eri-1(mg366)* animals relative to N2 controls (Figure 2.3A; see Supplemental Experimental Materials and Methods for target selection criteria). Consistent with the temporal expression pattern of class I sperm 26G RNAs, target silencing occurs in a relatively narrow window that corresponds to spermatogenesis through young adulthood (Figure 2.3A; Figure 2.8). By comparison, although class II oocyte/embryo 26G RNA levels steadily decline during larval development, their silencing effects persist throughout development (Figure 2.3A).

We next asked if the *eri-1*-dependent regulation of 26G RNA targets could be observed at the whole-transcriptome level. Using previously reported whole-genome microarray data that compared transcript expression profiles of L4 stage *eri-1* and N2 worms (102), we found that predicted targets of 26G RNAs are significantly up-regulated in the *eri-1(mg366)* mutant ($p<0.0001$, t-test) (Figure 2.3C). Conversely, genes up-regulated in the *eri-1* mutant background also were 9 fold enriched for 26G RNA targets (Supplemental Computational Methods). Taken together, the highly correlated expression patterns between 26G RNAs and their putative targets at the whole-transcriptome level further support the hypothesis that 26G RNAs directly regulate target gene expression in an *eri-1*-dependent manner.

To determine if target de-repression in *eri-1(mg366)* results in misexpression of target mRNAs in inappropriate tissues, we performed RNA *in situ* hybridization for select, relatively abundant (101) targets (*C04G2.8* and *ssp-16*) in dissected gonads. While the class I sperm 26G RNA targets *C04G2.8* and *spp-16* are up-regulated in the *eri-1* mutant (Figure 2.3D), they exhibit similar expression patterns in the male spermatogenic gonads of both the *him-8* and *eri-1; him-8* strains (Figure 2.3E). Thus, target de-silencing by class I sperm 26G RNAs in the *eri-1* mutant remains restricted to the male gonad and does not result in inappropriate, ectopic expression in either the male gonads or in the oogenic gonads of *eri-1* hermaphrodite animals (Figure 2.3E), indicating that 26G RNAs repress target expression in their cognate cell types.

2.3.4 Genetic requirements for 26G RNA biogenesis and function

Because small RNAs that start with a guanine nucleotide are thought to be products of an RdRP (103), we asked if RdRPs could play a role in biogenesis of 26G RNAs. The *C. elegans* genome encodes four RdRPs (*rrf-1*, *2*, *3*, and *ego-1*) (104). We examined 26G RNA expression in mutants for three viable RdRPs, *rrf-1(pk1419)*, *rrf-2(ok210)*, and *rrf-3(pk1426*). Since mutations in *ego-1* result in lethality (105), we used RNAi to deplete the *ego-1* transcript from N2 animals. While *rrf-1(pk1419)*, *rrf-2(ok210)*, and *ego-1(RNAi)* express normal levels of 26G RNAs, both classes of 26G RNAs are abolished in *rrf-3(pk1426*), which results in significant up-regulation of both classes of targets (Figure 2.4A; Figure 2.9). However, we note that RNAi-inactivation of *ego-1* does not completely abolish

ego-1 expression and therefore we cannot definitively conclude that the 26G RNAs are strictly *ego-1*-independent. If 26G RNAs are *bona fide* RdRP products, then transcripts they target should serve as templates for 26G RNA production. We determined that the expression of a 26G RNA (26G-S4) is dramatically reduced when its target *deps-1* is mutated and degraded by nonsense-mediated decay (see Supplemental Results and Figure 2.10). Interestingly, although 26G RNAs require the RRF-3 RdRP, they appear to possess a 5' monophosphate, as opposed to the 5' triphosphate of other known RdRP products (see Supplemental Results and Figure 2.11). In addition, with notable exceptions (21U RNAs), the presence of a 5' monophosphate on a small RNA is a signature of Dicer processing. Expression analysis in the *dcr-1(ok247)* mutant indicates that the 26G RNAs likely require DCR-1 for biogenesis (see Supplemental Results and Figure 2.12).

The non-overlapping identities of the two classes of 26G RNAs and the disparate targets they regulate suggested that they might be sorted into distinct RISCs. Argonaute proteins are central components of RISC and possess two conserved domains, PAZ and PIWI. Argonautes directly bind small RNAs (via both domains) and may possess target cleavage ("slicer") activity via the PIWI domain (106). *C. elegans* encodes 27 potential Argonautes with diverse functions, several of which have been found to be enriched during spermatogenesis or oogenesis (101, 107). We found that an Argonaute encoded by *ergo-1 (107)*, whose transcript is enriched during oogenesis (101), is required for the expression of class II oocyte/embryo 26G RNAs, but not for class I sperm

26G RNAs (Figure 2.4B). Consistent with this finding, only targets of class II oocyte/embryo 26G RNAs were up-regulated in the *ergo-1(tm1860)* mutant (Figure 2.9). The expression of two Argonautes, T22B3.2 and its close paralog, ZK757.3 (93.1% amino acid sequence identity), are enriched during spermatogenesis (101). Although the single mutant of either *t22b3.2(tm1155)* or *zk757.3(tm1184)* maintains wild-type expression levels of both classes of 26G RNAs, mutations in both *T22B3.2* and *ZK757.3* abrogate the expression of class I sperm 26G RNAs, but not class II oocyte/embryo 26G RNAs (Figure 2.4B). Similarly, only targets of class I sperm 26G RNAs are de-repressed in the double mutant (Figure 2.9). ERGO-1, T22B3.2, and ZK757.3 all possess the Asp-Asp-His catalytic "slicer" motif (107, 108), suggesting that they are capable of directly mediating endonucleolytic cleavage of their targets. Taken together, our data suggest that distinct RISCs guide the class I and class II 26G RNAs to their cognate targets for silencing.

What are the biological functions of 26G RNA-mediated target regulation? *eri-1* and *rrf-3* mutants, which lack both class I and class II 26G RNAs, are temperature-sensitive (*ts*) sterile due to defective spermatogenesis (100, 109). While the single Argonaute mutants of *T22B3.2* and *ZK757.3* exhibit near-wild-type levels of fertility, the double mutant, which is specifically defective in the expression of class I sperm 26G RNAs, is completely sterile at 25°C and can be fully rescued by crossing with WT males (Figure 2.13A-C). In contrast, the *ergo-1* Argonaute mutant, which is defective in the expression of class II oocyte/embryo 26G RNAs, displays near wild-type fertility. These findings

suggest that class I sperm 26G RNAs play an essential gene regulatory role during spermatogenesis. Loss of class II oocyte/embryo 26G RNAs does not result in any overt developmental phenotypes, as we did not observe any somatic defects in the *eri-1*, *rrf-3*, or *ergo-1* mutant. This is consistent with the finding that endo-siRNAs recently identified in fly soma and mouse oocytes appear to be dispensable for viability and reproduction (12, 86-89). Interestingly, mutants of *eri-1*, *rrf-3*, and *ergo-1* all exhibit an enhanced response to exogenous RNAi (100, 107, 109), whereas the *t22b3.2; zk757.3* double mutant does not (Figure 2.13E), suggesting that class II 26G RNAs may compete with the exogenous RNAi pathway for limiting common factors (100, 107, 109).

2.4 Discussion

In this study, we characterized a class of germline-enriched endo-siRNAs that are generated by a template-dependent mechanism and require the RRF-3 RdRP and the ERI-1 exonuclease for their biogenesis. In our model, class I and class II 26G RNAs are sorted into distinct, gamete-specific RISCs during germline development and differentially target discrete classes of target genes (Figure 2.4C). Class I 26G RNAs repress their target genes during spermatogenesis and mutations that abrogate their expression lead to male sterility. Class II 26G RNAs are maternally loaded and appear to be responsible for the clearance of maternal transcripts during zygotic development. In zebrafish, miR-340 clears hundreds of maternal mRNAs during the maternal-zygotic transition (21). In our model, the class II 26G RNAs not only begin to

clear a subset of target maternal mRNAs that are deposited, but also act to ensure that the maternal load of mRNAs continues to be cleared during filial development. The fact that the loss of class II 26G RNAs leads to enhanced RNAi phenotypes suggests that ongoing transcript clearance competes with exogenous RNAi for limiting factors.

In exogenous RNAi, primary siRNAs derived from Dicer processing of an exogenous dsRNA trigger initiate unprimed synthesis of secondary siRNAs mediated by RdRPs (110, 111). Unlike the primary siRNAs that possess a 5' monophosphate, these secondary siRNAs contain a 5' triphosphate modification. Similarly, we speculate that 26G endo-siRNAs might function as 5' monophosphorylated primary endo-siRNAs, whose biogenesis is likely *dcr-1*-dependent, to guide target cleavage and initiate the production of ~21nt secondary endo-siRNAs that further silence the 26G RNA targets. Interestingly, these ~21nt putative secondary siRNAs also appear to be 5' triphosphorylated (Figure 2.11) and, therefore, would be underrepresented in our deep sequencing datasets that enrich for RNAs possessing a 5' monophosphate group.

Our study raises other interesting questions. Why are certain genes targeted by 26G RNAs? How do ERI-1 and RRF-3 participate in the biogenesis of 26G RNAs? Why do the loss of sperm 26G RNAs and consequent up-regulation of targets lead to *ts* sterility? Further genetic and biochemical analysis may reveal additional factors and mechanisms that mediate the biogenesis, sorting, differential stability, target silencing, and developmental functions of the class I and class II 26G RNAs.

2.5 Materials and methods

Strains and sperm, oocyte, and embryo purifications. The Bristol N2 was used as the reference wild type strain. Mutant alleles used in this study include: LG I: *glp-4(bn2), fer-1(hc1), rrf-1(pk1417), rrf-2(ok210), deps-1(bn121), deps-1(bn124), smg-1(r861)*; LG II: *rrf-3(pk1426)*; LG III: *zk757.3(tm1184)*, *dcr-1(ok247)*; LG IV: *him-8(e1489), eri-1(mg366), t22b3.2(tm1155)*; LG V: *ergo-1(tm1860)*. Sperm and oocytes from *him-8(e1489)* and *fer-1(hc1)*, respectively, were purified as described (112) with some modifications. See Supplemental Experimental Materials and Methods for additional details.

RNA analysis. Total RNA isolation was carried out using TriReagent (Ambion) following the vendor's protocol. 5' monophosphate-bearing small RNA libraries were constructed as described (113). Due to limitation in sensitivity, relatively abundant 26G RNAs were selected for northern blotting as described (114) using 5-10 µg of total RNA and Starfire DNA probes (IDT). For RT-qPCR analysis of small RNAs, custom small RNA Taqman assays (Applied Biosystems) were performed following the vendor's protocol. For quantification of mRNAs, 250ng-1µg of total RNAs were converted into cDNAs with Multiscribe Reverse Transcriptase (Applied Biosystems) following the vendor's protocol. See Supplemental Experimental Materials and Methods for additional details.

Germline RNA in situ hybridization. RNA *in situ* hybridization was performed with dissected gonads according to Lee and Schedl (115). Antisense

cDNA fragments labeled with DIG DNA labeling Mix (Roche) for *C4G2.8* (547bp) and *ssp-16* (102bp) were used as probes.

RNA interference. Gene inactivation by RNAi was performed as described (116) using clones from the Ahringer RNAi library.

Computational methods. See Supplemental Computational Methods.

2.6 Acknowledgements

The authors thank John Moran, Harrison Gabel, Chi Zhang, Tammy Wu, Kris Gunsalus, Scott Kennedy, Patrick Hu, and Allison Billi for helpful comments. We thank Sylvia Fischer for *dcr-1(ok247)* total RNA, the *Caenorhabditis* Genetics Center and Shohei Mitani for strains, Marco Marra and Martin Hirst of the British Columbia Cancer Centre for Solexa deep sequencing, and David Miller of Applied Biosystems for Taqman probes. This research was supported by an NSF CAREER Award, an NIH S06 GM52588 grant to D.S.C., and an NIH HG004276-01 grant to J.K.K.

2.7 Supplemental Results

2.7.1 Requirement of target mRNA transcript for 26G RNA biogenesis.

deps-1 is a gene whose 3'UTR appears to be targeted by a class I sperm 26G RNA (26G-S4) (Figure 2.10A). Two alleles of *deps-1* (*bn121* and *bn124*) introduce premature stop codons into the gene and destabilize *deps-1* transcripts (Figure 2.10A) (117). In both alleles, the expression of 26G-S4 is significantly

depleted (>10-fold), while expression of other 26G RNAs that do not target *deps-1* (26G-S5, -S6) is not affected, supporting the requirement of *deps-1* transcript as a template for 26G-S4 production (Figure 2.10B). We attempted to rescue 26G-S4 expression by crossing the *deps-1* mutants into the *smg-1(r861)* background, which stabilizes transcripts with premature stop codons that are degraded by the nonsense-mediated mRNA decay pathway (118). While the expression of *deps-1* mRNA in the double mutants is still below wild-type levels, we observed a noticeable increase in 26G-S4 expression in one (*deps-1(bn121); smg-1(r861)*), but not the other (*deps-1(bn124); smg-1(r861)*), double mutant (Figure 2.10D); this is likely because the expression of *deps-1* mRNA remains below a threshold level for 26G-S4 synthesis or detection in the *deps-1(bn124); smg-1(r681)* double mutant (Figure 2.10C).

2.7.2 5' monophosphorylation of 26G RNAs

In *C. elegans*, during exogenous RNAi, a similar RdRP-mediated process programmed by *rrf-1* generates secondary siRNAs to amplify the silencing signal (110, 111, 119). These secondary siRNAs start with a guanine nucleotide and are triphosphorylated at 5' end (5'-PPP). However, although 26G RNAs require the RRF-3 RdRP, they are suitable substrates for T4 RNA ligase-mediated 5' linker ligation (Figure 2.11), suggesting that most 26G RNAs possess a 5' monophosphate group (110, 111). These findings further support our original small RNA cloning procedure that identified the 26G RNAs based on a ligation reaction that selected small RNAs containing a 5' monophosphate group.

2.7.3 The biogenesis of 26G RNAs is likely Dicer-dependent.

The presence of a 5' monophosphate on a small RNA is a signature for Dicer processing with notable exceptions (*e.g.* 21U RNAs) (39). Because 26G RNAs appear to be 5' monophosphorylated, we asked if they are processed by Dicer. *C. elegans* encodes a single Dicer ribonuclease, *dcr-1*, which is essential for germline development and viability (92-94). Homozygous *dcr-1(ok247)* null animals that are produced by *dcr-1*/+ heterozygotes live until adulthood but exhibit pleiotropic defects including complete sterility, abnormal vulval structures, and unfertilized oocytes (92) (Figure 2.12A). At the young adult stage, levels of both microRNAs (*let-7* and miR-1) and 26G RNAs (26G-O1, -O2, -O3, and -194) significantly decrease in the *dcr-1* null mutant, relative to the *dcr-1* heterozygous animals (Figure 2.12C). The degree to which the expression of microRNAs and 26G RNAs is compromised correlates with the severity of *dcr-1* phenotypes (Figure 2.12A) and with the level of maternal *dcr-1* mRNA remaining in the *dcr-1* mutant (Figure 2.12B). Thus, *dcr-1* null animals with the severe phenotype of bursting exhibit the lowest level of maternal *dcr-1* mRNA, microRNAs, and 26G RNAs. Surprisingly, the expression of 21U RNAs (21UR-342 and 21UR-684), which was reported to be *dcr-1*-independent (39), also decreases to a similar degree as the microRNAs and 26G RNAs in the young adult *dcr-1* mutant (Figure 2.12C). The germline itself is required for the biogenesis of 26G and 21U RNAs (Figure 2.1E), but dispensable for the somatic expression of the microRNAs *let-7* and miR-1. Therefore, we next asked if we could discriminate *dcr-1*-dependence

among the three types of small RNAs during an earlier period of germline development, the L4 larval stage, when the cumulative effects of the *dcr-1* mutation on germline development are less severe (Figure 2.12D). In agreement with the results in the young-adult stage, the *dcr-1* homozygotes at L4 stage again exhibited decreased expression of microRNAs (*let-7* and miR-1) and 26G RNAs (26G-S4, -S5, and -S6). However, the expression of 21U RNAs (21UR-342 and 21UR-684) does not decrease in the L4-stage *dcr-1* homozygotes, supporting previous observations of *dcr-1*-independence (39); on the contrary, their levels appear to increase. Taken together, our data suggest that the biogenesis of 26G RNAs likely requires *dcr-1.* However, because the 26G RNAs are generated in the germline, we cannot conclusively rule out the possibility that the decrease in 26G RNA expression may be an indirect consequence of defects in germline development exhibited by the *dcr-1* (-/-) mutant.

2.8 Supplemental Computational Methods

Sequence processing. All raw sequences (consolidating both 454 and Solexa) were processed with a custom Perl script to remove linker sequences and then mapped against the WS190 *C. elegans* genome using BLAST (120). Sequences matching the genome with 0-2 mismatches were retained. Reads not matching the genome were mapped against Expressed Sequence Tags (EST) using BLAST to identify sequences that span exon-exon junctions. For reads matching more than one genomic locus, counts were normalized according to Ruby *et al.* (37). For example, if a sequence had 20 reads and matched 2

genomic loci, each locus was assigned 10 reads. For all endo-siRNA analyses, reads corresponding to microRNAs (37), 21U RNAs (39, 96), and putative degradation products of non-coding RNAs (*i.e.* rRNAs, tRNAs, snRNAs, snoRNAs) were identified and excluded.

Genomic mapping of 26G RNAs. As outlined in Figure 2.5, we applied sequential filters to retain 26G RNAs with ≥ 2 reads in the 11 sequenced libraries and mapped them sequentially to WormBase (WS190) and predicted gene models (Twinscan and Genefinder in WS190). Because 3'UTR regions are not well annotated, reads immediately downstream (within 500bp) of stop codons were annotated as overlapping with 3'UTR, which agrees well with the distribution of known 3' UTR lengths of annotated genes in WormBase (Figure 2.6). The remaining intergenic 26G RNA sequences (23.3%) may also target genes yet to be identified by WormBase gene annotations and predictions.

Cluster analysis of 26G RNAs. 26G RNAs (≥ 2 total reads) were clustered using Cluster 3.0 software (copyright Stanford University, 1998-99) and visualized using Java TreeView (open source). Clusters of the class I sperm 26G RNAs and the class II oocyte/embryo 26G RNAs were extracted from Java TreeView.

Target analysis of 26G RNAs. Targets of class I sperm 26G RNAs and class II oocyte/embryo 26G RNAs (extracted from clustering analysis) were annotated as spermatogenesis-enriched, oogenesis-enriched, germline-intrinsic, and "others" according to Reinke *et al.* (101). For microarray analyses, raw CEL data from Asikainen *et al.* (102) were downloaded from NCBI Gene Expression

Omnibus (Series GSE8659) and processed with dChip software (121). Probe intensities corresponding to targets of sperm 26G RNAs were extracted from the CEL data. The 9 fold enrichment described in main text was derived as follows: out of 72 genes that are significantly upregulated in *eri-1* (102), 36 (50%) are 26G RNA targets, representing a 9 fold enrichment, as 1118 genes (5.5% of all protein coding genes) are targeted by 26G RNAs.

2.9 Supplemental Experimental Materials and Methods

Sperm, oocyte, and embryo purifications. For sperm isolation, we used the *him-8(e1489)* strain, which increases the percentage of XO males to ~37% of the population versus ~0.2% males in the N2 wild-type strain (122). Male worms from the *him-8(e1489)* strain were further isolated from hermaphrodites by filtering through a 35 μm nylon mesh filter as described (112), resulting in >95% males in the final sample. Isolated *him-8(e1489)* males were then subjected to 20,000 psi for 1 min, 3 times, to extrude and increase the yield of purified sperm. We used the *fer-1(hc1)* strain, which produces nonfunctional sperm at 25°C (123), to obtain purified unfertilized oocytes. The *fer-1(hc1)* worms grown at 25°C were disrupted briefly in a Waring blender to release more oocytes from the body cavity. Sample purity (>95%) was inspected by DAPI staining and microscopy. Isolation of embryos from gravid adult worms was performed as described (124).

Construction of small RNA sequencing library. RNA oligos were purchased from Dharmacon and DNA oligos from Integrated DNA Technologies. Six Solexa libraries were constructed and sequenced on the 1G Genome

Analyzer (Solexa/Illumina): N2 (mixed stage), sperm, oocyte, embryo, *eri-1(mg366)*, and *glp-4(bn2)* young adults (YA). Five 454 libraries (sperm, oocyte, N2, *eri-1(mg366)*, and *glp-4*) were sequenced on the Genome Sequencer FLX system (454/Roche).

RT-qPCR analysis of small RNA and mRNA levels. Custom small RNA Taqman assays were designed and synthesized by Applied Biosystems (125). For each reaction, 50ng of total RNA was converted into cDNA with Multiscribe Reverse Transcriptase (Applied Biosystems) following the vendor's protocol. The resulting cDNAs were analyzed by a Realplex^2 thermocycler (Eppendorf) with TaqMan Universal PCR Master Mix, No AmpErase UNG (Applied Biosystems). Relative expression levels of small RNAs were calculated based on 2^{-ct} method (126). For oocyte/embryo 26G RNA quantifications, miR-35 was used for normalization. For sperm 26G RNA quantifications, miR-1 was used for normalization. Gene targets of each class of 26G RNAs were selected based on 26G RNA cluster analysis (described below in supplementary computational methods). For quantification of mRNA levels, 250ng-1µg of total RNAs was converted into cDNAs with Multiscribe Reverse Transcriptase (Applied Biosystems) following the vendor's protocol. cDNAs were analyzed by a Realplex^2 thermocycler (Eppendorf) using Power Sybr Green PCR master mix (Applied Biosystems). Relative mRNA levels were calculated based on 2^{-ct} method using *act-1* for normalization.

Table 2.1 Oligos for RT-qPCR.

Gene	Forward (5' to 3')	Reverse (5' to 3')
act-1	CCAGGAATTGCTGATCGTATGCAGAA	TGGAGAGGGAAGCGAGGATAGA
C04G2.8	CGTGCTTCGACTGCAAAGAAGA	TTCTGTTGGCTTCTGCTGCG
C32E8.4	GAGCAACTTCTGCCGAAGGAA	CTTCAGGTTCTCCTTGAGCG
C40A11.10	AATGGCTCCTTGAAAAGATCG	TACATTTCCGCCACGTTGAAA
deps-1	GAAGGCTATGGCCGAAGTTCG	CAATGCGGTAACGGACAGATTT
dlc-6	CCGAAGGTTAAGCCACGTCATT	CTGCCATTGTGTATCATAATCCG
E01G4.7	GCACAAGGTTTCGTTCTTGGTG	AGTGACATCCCTTCTGATCG
F39E9.7	CCCAGTGGCCCAATTAAACG	CCCACGGCTTGTTCTTTGACA
F43E2.6	TGTAGGCGACGAGACTGATCG	TGCCGATGTTTCTGAGATGTCTT
F55B11.1	TTGATCGAGTCTCACTTTCCG	AAAGTCCACTGGTTCGTGATGAAT
F55C9.5	ACCATTGGAGCACGTAAATCAA	GGTCCTAATAATAAAGTTGCGTCG
fbxa-65	ACTTACAAGGATCAAGAAAAGCG	CCTTGACCGCTATTCCGAGAAA
fbxb-37	ATCGAAAGATGGAATACAAACCG	GACAAACATCCATCACATTCTTCG
gska-3	CGAGCAGACGACTCTGTGGAA	TTATTGAAACGCACAGTCTTCTCG
iff-1	CGAAGACCATAGAGAGTATGTCCG	CGAGCATTGCTTCGGGAAAGTA
K02E2.6	CAGTGGTACAAGTGGGAGTAAACG	AATTGGCAAGTAACTGATTCCG
K03H1.12	CAAAATTGCCACTTGTGATTCG	TCCAGTGAAGAGTGTCAAGAACCA
msp-49	ATTAACTCCTCGGCTCGCCG	AGCTTCCTTTGGGTCGAGGAC
snf-6	GGATTGTTGGCTACTGGCCG	TCAAGCCAAAGGAAGCAAAGAA
sod-1	GATCTATGGTTGTTCATGCCG	CTTCTGCCTTGTCTCCGACTCC
ssp-16	GTCATCAAACAACAATGAGTACCG	GCTCCAGCAGTGCGAGTGAT
ssp-19	GCACCGAAGGAAGACAAGCTG	GAGCCACTGCAACAAAAGCG
T05E12.8	TTCCATTTGAGGATTTTGCTACG	ATTATTTGGATGGCAGCCGATG
T08B2.12	GAAACCAATGCTCCAGTTGATAC	GATGAAAGCGATGGACGAGAAG
T25G12.11	ACGTGCTTTCTGATTCACTCCG	CATGGGTGGGATGAGAGCAC
tax-2	GATTAATCCAAGACAAGTTCCTAAATTGAT	TTCAATTCTTGAACTCCTTTGTTTTTC
Tc1	AACCGTTAAGCATGGAGGTG	CACATGACGACGTTGAAACC
Tc3	GAGCGTTCACGGAGAAGAAG	AATAGTCGCGGGTTGAGTTG
tdc-1	GAACTTCGTCAGAGATTCCCG	TCTCAACGGAAGAATGGGCTTC
U6	TGGAACAATACAGAGAAGATTAGCA	CTTCACGAATTTGCGTGTCAT
W05H12.2	GCTCAAGACCAGATAATGCTTGGA	CAATCCCAAAGATTCAATACCG
Y37E11B.2	AATGGAGACTCTTCTTCCACCCG	AGCGAAGGCATTGATCTTGGTT
Y7A5A.11	CCATTACTTTCAACATGCCG	TCCTTGTTCCAGCACTAGCAGA
Y82E9BR.20	CTCCCGCTTTCTTGATGTATTG	AGTCCGAACTCATCCAAAGCAG
ZC168.6	GTCCAGTTTATGGGTTCGTGGATG	AGTCTCTTCGGCTGGCACTTC
ZC328.1	GGGCGGTCATTTCTATTGTTTG	GCCAAATTGGTCCGTAATCTTGT
ZK484.5	CCGTCAGACAACTGCTCTCCTC	GGTTGGGCTGCTTCAGAGTC

Table 2.2 Oligos for small RNA cloning.

5' RNA adaptor:	5' GUUCAGAGUUCUACAGUCCGACGAUC 3'
3' RNA adaptor:	5' pUCGUAUGCCGUCUUCUGCUUGidT 3' p = phosphate; idT = inverted deoxythymidine
RT-primer (DNA):	5' CAAGCAGAAGACGGCATACGA 3'
P7 primer (DNA):	5' CAAGCAGAAGACGGCATACGA 3'
P5 long primer (DNA):	5' AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA 3'

Table 2.3 Oligos for northern blotting.

21UR-1	5' GCACGGTTAACGTACGTACCA /3StarFire/ 3'
26G-O1	5' TTGAAAATAATCTACCGTTTCTGAGC /3StarFire/ 3'
26G-O2	5' CATTTGCTGCAATTATGAGTCATAAC /3StarFire/ 3'
26G-O3	5' AAAAGTATCCGACTTTCGAGTTTGTC /3StarFire/ 3'
26G-O5	5' CCCCTCTTTTCTTCTGCATTCCCATC /3StarFire/ 3'
26G-O6	5' ATGAAATGCCAGATGAATCCTTCTAC /3StarFire/ 3'
26G-S1	5' AATTATGTATTCTCGTCCTCCATAGC /3StarFire/ 3'
26G-S5	5' TACCATGTCGCTCACTGCTGATCCAC /3StarFire/ 3'
cel-miR-35	5' ACTGCTAGTTTCCACCCGGTGA /3StarFire/ 3'
cel-miR-1	5' TACATACTTCTTTACATTCCA /3StarFire/ 3'

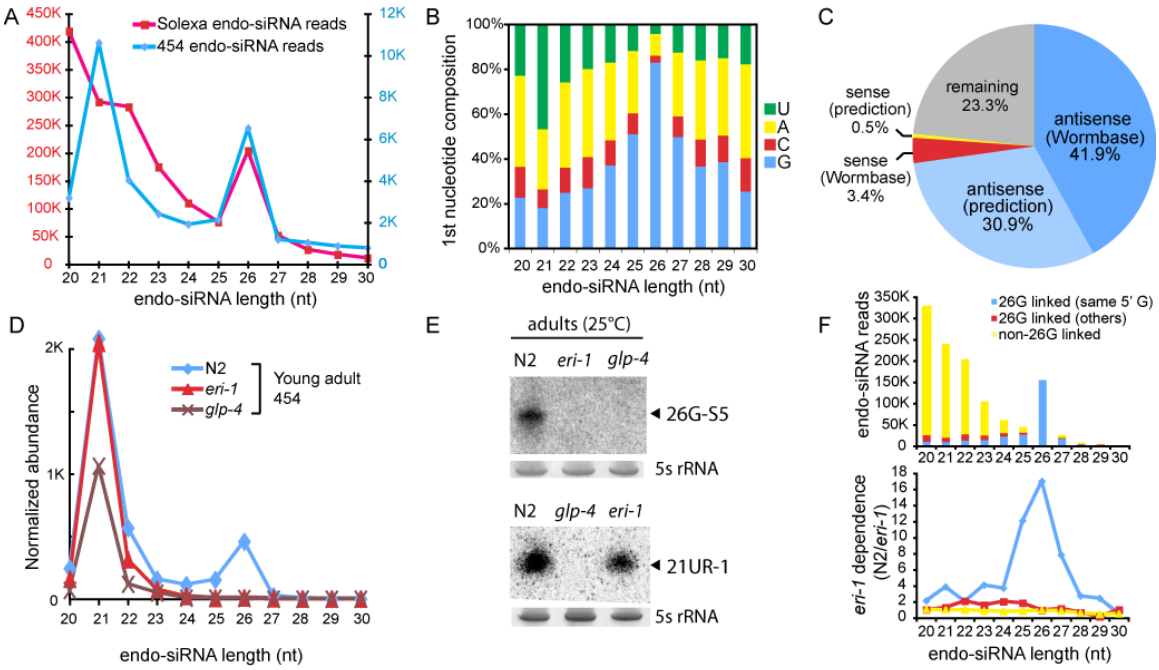


Figure 2.1 26G RNAs are germline-enriched endogenous siRNAs.
A.) Length distribution of endo-siRNAs exhibits a bimodal pattern, peaking at both 21nt and 26nt length. Small RNA libraries of mixed-stage N2 animals and purified male sperm (*him-8(e1489)*), oocytes (*fer-1(hc1)*), and N2 embryos were sequenced by Solexa (Illumina); libraries of N2 young adults, sperm, and oocytes were sequenced by 454 (Roche). The *him-8(e1489)* mutation increases the percentage of XO males to ~37% of the population (122) versus ~0.2% males in the N2 wild type strain; the *fer-1(hc1)* mutation results in nonfunctional sperm at 25°C (123), enabling purification of unfertilized oocytes. **B.)** First nucleotide identity of endo-siRNAs. 26nt endo-siRNAs have a strong preference for guanine as the first nucleotide (83%). **C.)** The majority of 26G RNAs are antisense to known and predicted coding transcripts. **D.)** Normalized length distribution of endo-siRNAs in N2, *eri-1(mg366)*, and *glp-4(bn2)* young adult libraries sequenced by 454 (Roche). The abundance was normalized to 100K effective small RNA reads (excluding putative degradation products of abundant ncRNAs). **E.)** Northern blotting validates the lack of 26G RNA expression in *eri-1(mg366)* and *glp-4(bn2)* mutants. Total RNA from N2, *eri-1(mg366)*, and *glp-4(bn2)* adult worms was probed for a 26G RNA (26G-S5) and a 21U RNA (21UR-1). The expression of the germline-derived 21U RNA (21UR-1) is not *eri-1*-dependent. 5S rRNA serves as the loading control. **F.)** Endo-siRNAs were classified as 26G RNA-linked (targeting the same genes) or non-26G RNA-linked (targeting other genes or intergenic regions). Most 26G RNA-linked endo-siRNAs start with the same 5' G. A small fraction of shorter (20-24nt) endo-siRNAs are 26G RNA-linked. The bottom panel plots the *eri-1* dependence as measured by the ratio of 26G RNA sequence counts in N2 vs. *eri-1*.

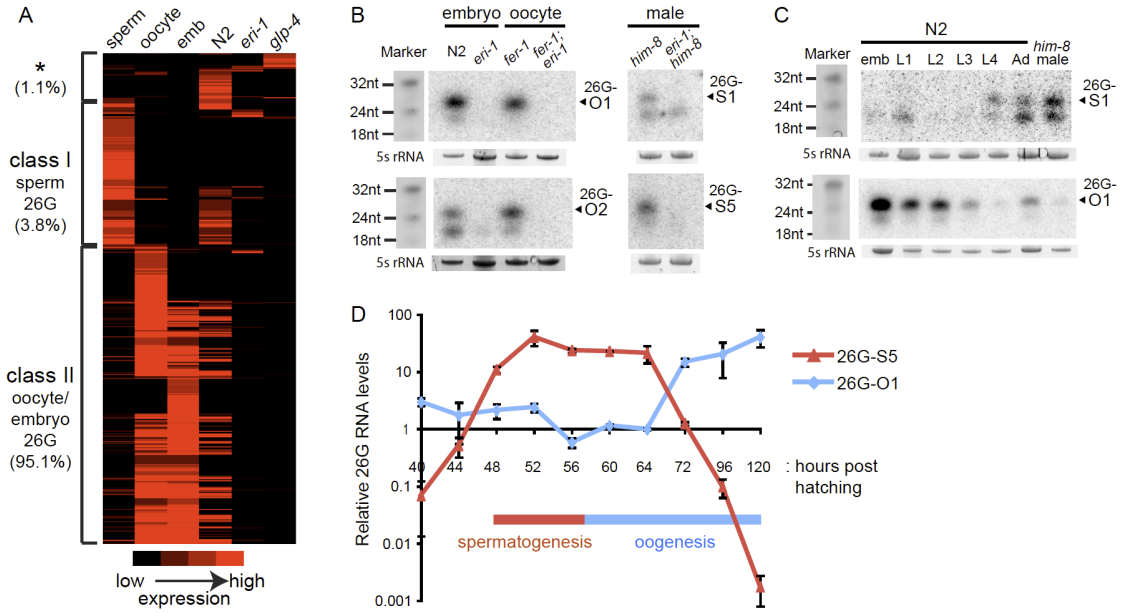


Figure 2.2 Two classes of 26G RNAs exhibit different expression patterns. **A.**) Hierarchical clustering reveals two major classes of 26G RNAs: class I sperm 26G RNAs (3.8% of total reads) and class II oocyte/embryo 26G RNAs (95.1%). 26G RNA reads matching to the genome with at least two counts were included in the analysis (4,002 unique sequences; 156,204 total reads). The asterisk (*) indicates a small fraction (1.1%) of 26G RNA sequences that do not fall into either class I or II categories. **B.)** Both classes of 26G RNAs are dependent on *eri-1* for their expression. Total RNA from N2 and *eri-1(mg366)* embryos and *fer-1(hc1)* and *fer-1(hc1); eri-1(mg366)* oocytes were probed for class II oocyte/embryo 26G RNAs (26G-O1, -O2). Total RNA from *him-8(e1489)* and *eri-1(mg366);him-8(e1489)* adult males was probed for class I sperm 26G RNAs (26G-S1, -S5). 5S rRNA serves as a loading control. **C.)** Class I and class II 26G RNAs are expressed in distinct periods during development. Total RNA from embryos, four larval stages, adult hermaphrodites, and *him-8(e1489)* adult males was analyzed by northern blotting with probes for a class I sperm 26G RNA (26G-S1) and a class II oocyte/embryo 26G RNA (26G-O1). Synthetic RNA oligos stained with EtBr serve as size markers, 5S rRNA as a loading control. **D.)** Analysis of 26G RNA levels during germline proliferation assayed by RT-qPCR. The expression of class I sperm 26G RNA (26G-S5) and class II oocyte/embryo 26G RNA (26G-O1) correlates with the time windows for spermatogenesis and oogenesis, respectively. The X-axis represents hours post-hatching at 20°C.

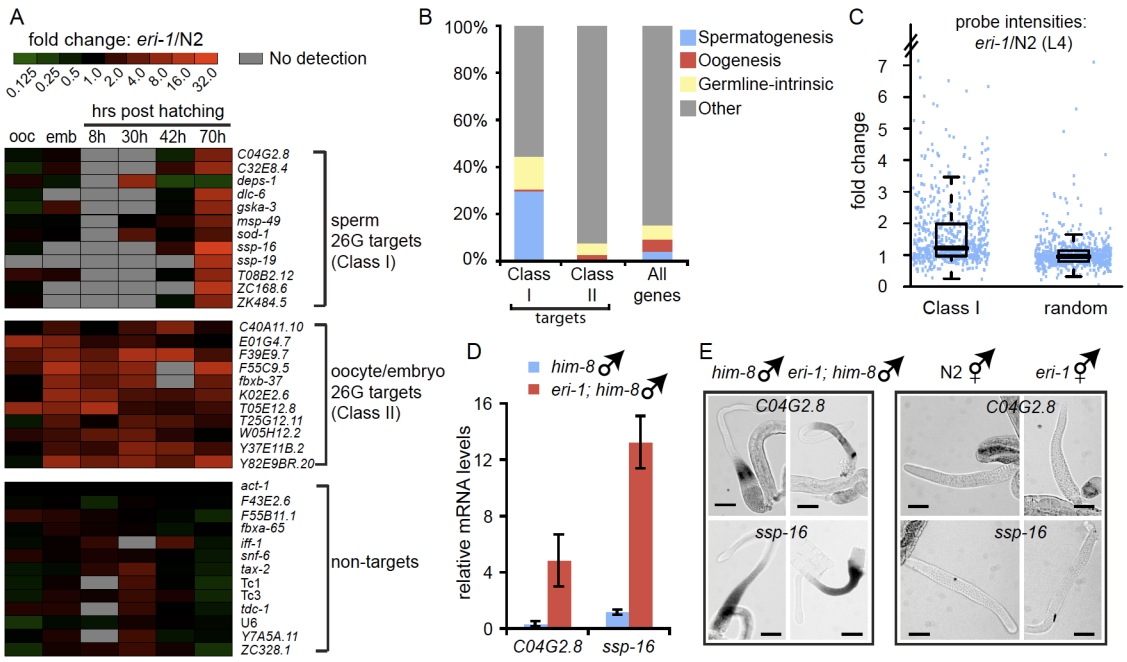


Figure 2.3 Two classes of 26G RNAs silence non-overlapping sets of mRNA transcripts.

A.) Gene targets of 26G RNAs are desilenced in the *eri-1(mg366)* background. Differential gene expression profiles between N2 and *eri-1(mg366)* for 12 targets of class I sperm 26G RNAs, 11 targets of class II oocyte/embryo 26G RNAs, and 13 non-targets were measured by RT-qPCR. The level of fold up-regulation is represented according to the red-green color scheme indicated in the top panel. **B.**) Gene class analyses of class I sperm and class II oocyte/embryo 26G RNAs. Targets of class I sperm 26G RNAs (573 genes) are significantly overrepresented in genes expressed during spermatogenesis, while targets of class II oocyte/embryo 26Gs (243 genes) are depleted of germline enriched genes. **C.**) Genes targeted by class I sperm 26G RNAs are up-regulated in the *eri-1(mg366)* mutant. Each point indicates the fold change in probe intensity corresponding to predicted targets of 26G RNAs (728 probes corresponding to 589 genes). Randomly selected probes do not show significant up-regulation in the *eri-1(mg366)* mutant. **D.)** Two sperm 26G RNA target mRNAs, *C04G2.8* and *ssp-16*, are up-regulated in *eri-1(mg366)*; *him-8(e1489)* males relative to *him-8(e1489)* males. mRNA levels were quantified by RT-qPCR and normalized to *act-1*. **E.**) Loss of 26G RNA expression does not induce inappropriate ectopic expression of targets. RNA *in situ* hybridization of dissected gonads was performed with probes for the class I sperm 26G RNA targets *C02G2.8* and *ssp-16*. In both wild type and *eri-1* backgrounds, expression of these two genes remained restricted to the spermatogenic gonad. No ectopic expression of the class I 26G RNA targets was observed in the hermaphrodite oogenic gonads. Bar, 50µm.

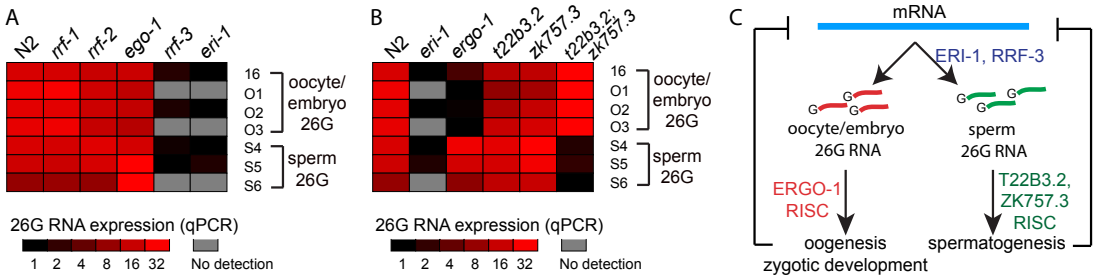


Figure 2.4 Genetic requirements for 26G biogenesis and function.
A.) RT-qPCR analysis of 26G RNA expression in *rrf-1(pk1417), rrf-2(ok210), rrf-3(pk1426*), and *ego-1(RNAi).* Mutation of *rrf-3* abrogates the expression of both sperm and oocyte/embryo 26G RNAs, while the 26G RNAs are expressed at wild-type levels in the mutants of *rrf-1* and *rrf-2,* as well as in RNAi-inactivation of *ego-1.* **B.)** An oogenesis-enriched Argonaute encoded by *ergo-1* is required for class II oocyte/embryo 26G RNA expression but dispensable for class I sperm 26G RNA expression. The *t22b3.2(tm1155)*; *zk757.3(tm1184)* double mutant is defective in sperm 26G RNAs but expresses normal levels of oocyte/embryo 26G RNAs. **C.**) Proposed model for 26G RNA biogenesis and function. See text for details.

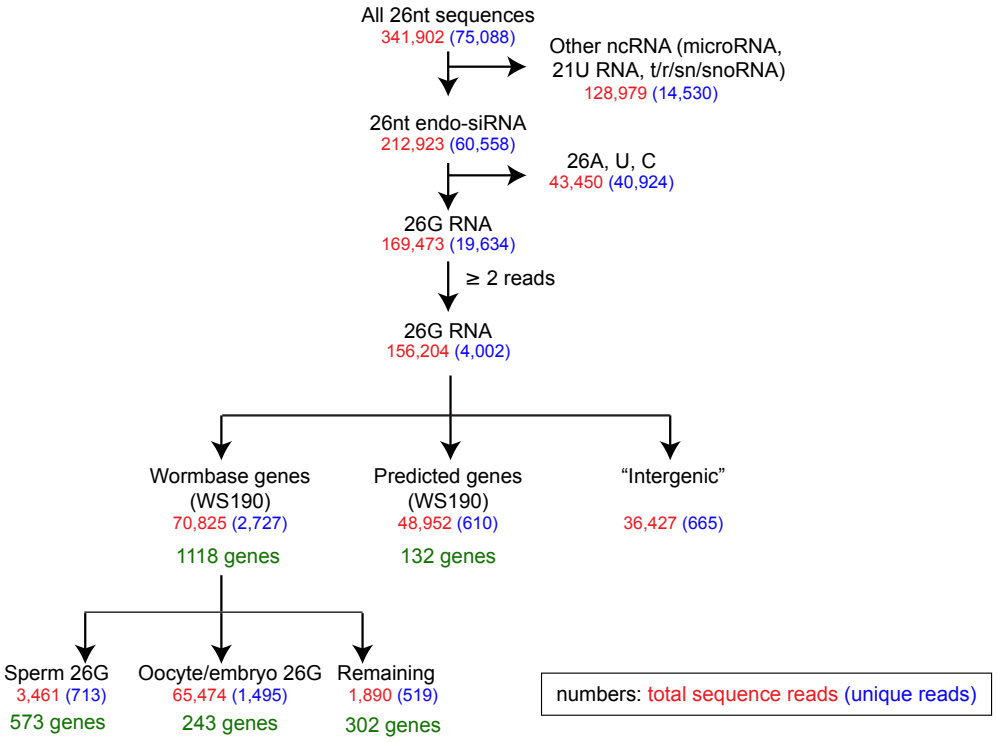


Figure 2.5 Computational pipeline for 26G RNA annotations.
All 26nt genome BLAST hits were extracted from our datasets. Sequences matching noncoding RNAs (i.e. tRNAs, rRNAs, snRNAs, snoRNAs) and other classes of small RNAs (microRNAs, 21U RNAs) were identified and excluded from the analyses. Two additional filters were applied to retain sequences starting with guanine and having ≥2 sequence reads. 26G RNAs mapping within 500bp downstream of WormBase gene annotations (WS190) and gene predictions (Twinscan, Genefinder predictions from WormBase) were sequentially annotated. In sum, 1,118 WormBase-annotated genes and 132 WormBase-predicted genes were identified to be targets of 26G RNAs. 26G RNAs derived from WormBase-annotated genes were further clustered into sperm 26G RNAs (with 573 gene targets) and oocyte/embryo 26G RNAs (with 243 gene targets).

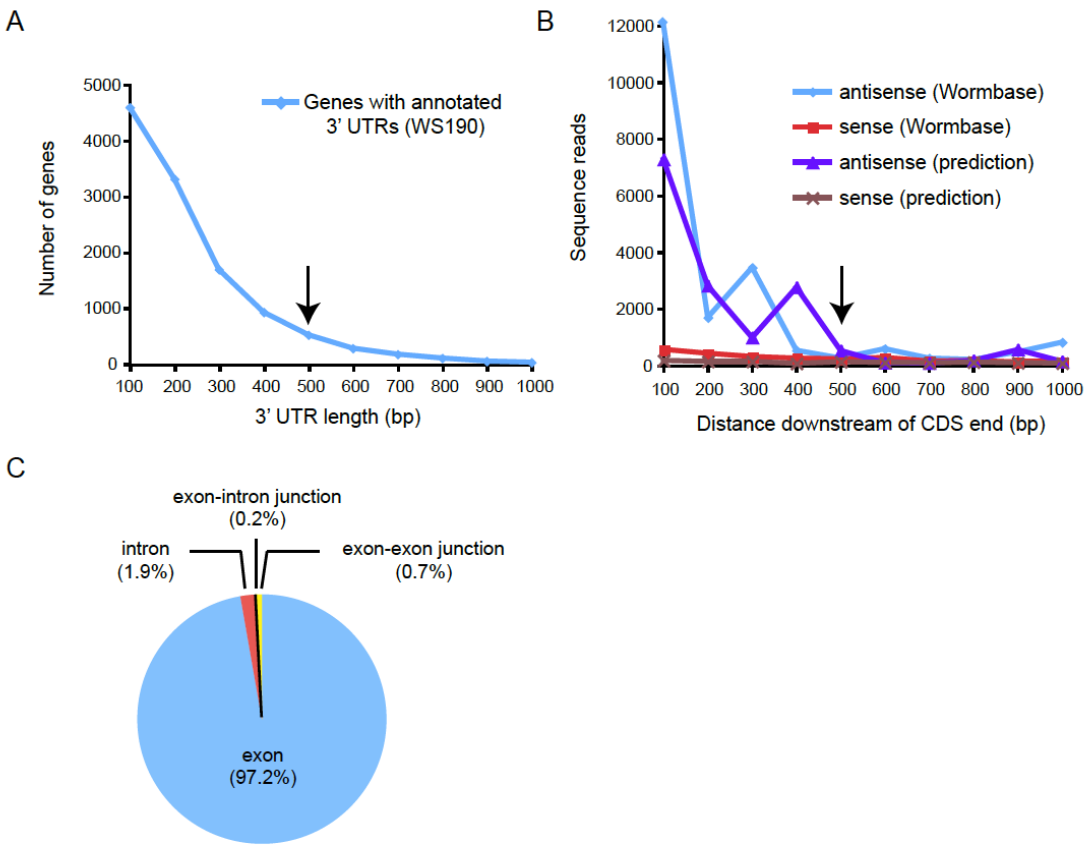


Figure 2.6 Distribution and mapping of 26G RNAs.
A.) The 3' UTR length distribution of genes in WormBase. Arrow at 500nt indicates the 95% cutoff.
B.) Number of 26G RNA reads that mapped within every 100bp up to 1Kb downstream of the ends of the coding sequences (stop codons) was plotted. The majority of reads are antisense to mRNAs and map within 500 bp (arrow) downstream of stop codons. **C.)** 26G RNAs mapping to exons and introns. 26G RNA counts matching exons, introns, exon-intron junctions and exon-exon junctions of WormBase genes were plotted. The majority of reads (97.9%) are derived from exons.

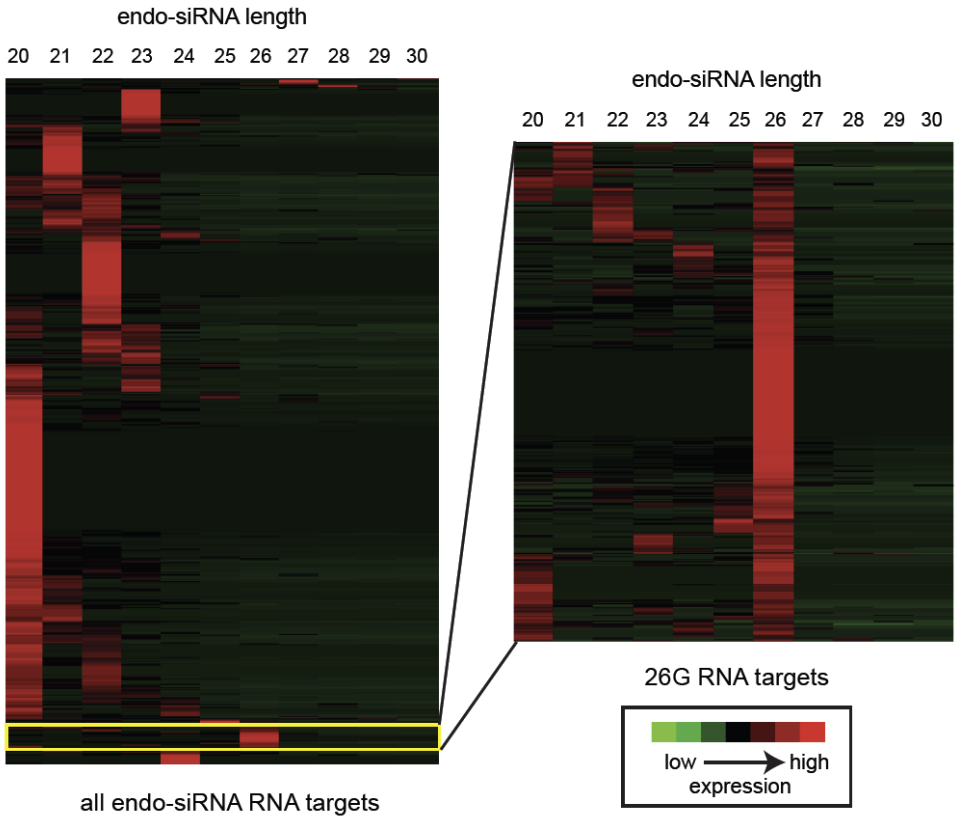


Figure 2.7 26G RNA targets are a unique class of genes.
Endo-siRNA targets (WormBase WS190) were clustered (left) based on the abundance of endo-siRNAs of different lengths. 26G RNA targets are predominantly targeted by 26G RNAs (right).

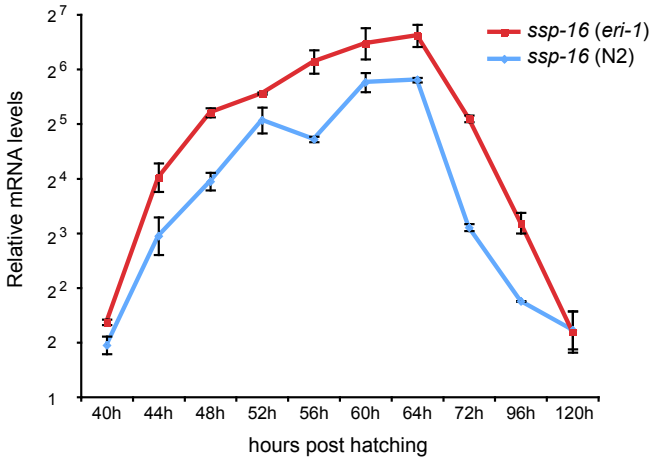


Figure 2.8 *ssp-16* (a target of sperm 26G RNA) is de-repressed starting from spermatogenesis until young adulthood in the *eri-1* mutant.
The X-axis represents hours post hatching at 20°C; the Y-axis indicates relative mRNA abundance in log2 scale. Relative mRNA levels were examined by RT-qPCR and normalized to *act-1*.

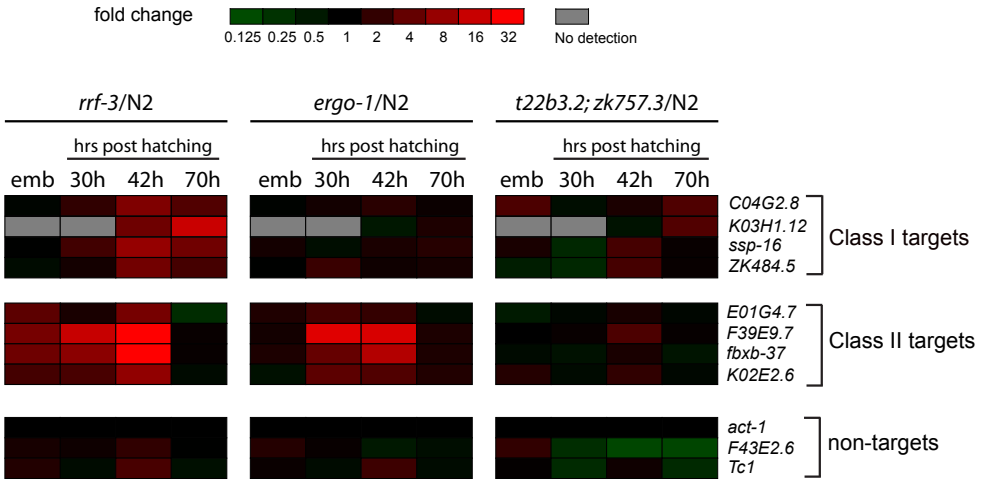


Figure 2.9 Differential gene expression profiles of 26G RNA targets in N2, *rrf-3*(*pk1426*), *ergo-1*(*tm1860*), and the *t22b3.2*(*tm1155*); *zk757.3*(*tm1184*) double mutant.

The transcript levels of 4 targets of class I sperm 26G RNAs, 4 targets of class II oocyte/embryo 26G RNAs, and 3 non-targets were examined. For example, the class I targets *C04G2.8* and *K03H1.12* are 3-fold up-regulated at 70hrs. Relative mRNA levels were examined by RT-qPCR and normalized to *act-1*. The fold up-regulation was represented according to the red-green color scheme shown (top panel).

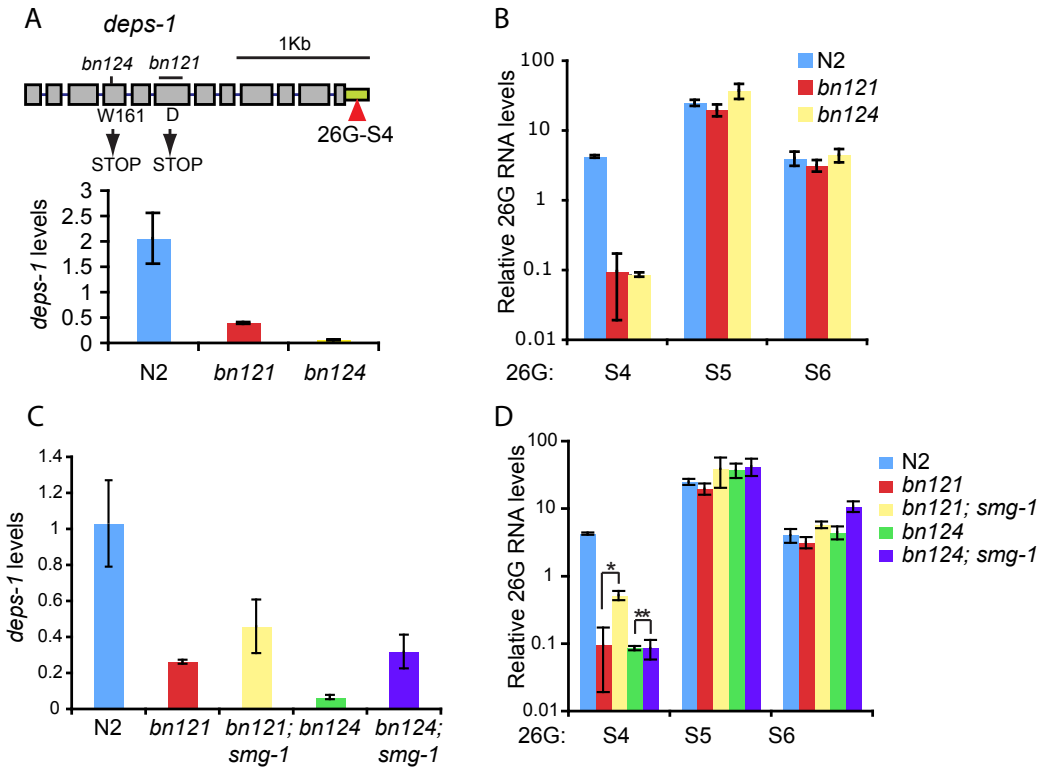


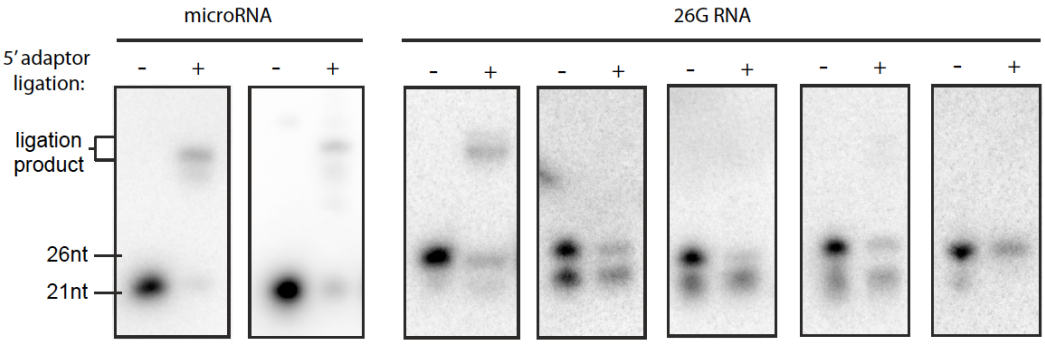
Figure 2.10 Requirement of target mRNA transcript for 26G RNA biogenesis.

A.) Two *deps-1* mutant alleles (*bn121* and *bn124*) harbor premature stop codons that destabilize the *deps-1* transcript. *deps-1* mRNA levels are measured by RT-qPCR and normalized to *act-1*. Error bars indicate standard deviation for replicates.

B.) The expression of the class I 26G RNA 26G-S4, which is antisense to the *deps-1* 3'UTR (green), is compromised in the *deps-1* mutants, while the expression of other sperm 26G RNAs that do not target *deps-1* (26G-S5 and -S6) remains unchanged. 26G RNA levels were measured by RT-qPCR and normalized to miR-1. Error bars indicate standard deviation for replicates.

C.) The expression of *deps-1* mRNA from the *deps-1* nonsense mutants (*bn121* and *bn124*) is partially restored in the nonsense decay mutant *smg-1*(*r861*), but still falls below WT levels.

D.) A noticeable increase of 26G-S4 levels, but not 26G-S5 and 26G-S6, is seen in the *deps-1*(*bn121*); *smg-1*(*rr861*) double mutant, relative to the *deps-1*(*bn121*) single mutant (*).



	miR-1	miR-35	26G-O1	26G-O2	26G-O3	26G-O5	26G-O6
26nt ratio (- : +)	NA	NA	1 : 0.32	1 : 0.26	1 : 0.23	1 : 0.26	1 : 0.57
~21nt ratio (- : +)	1 : 0.08	1 : 0.17	1 : 0.75	1 : 0.59	1 : 0.92	1 : 0.81	1 : 0.23

Figure 2.11 Depletion analysis indicates that 26G RNAs are suitable substrates for T4 RNA ligase-mediated ligation.
Small RNAs (18-32nt) were isolated by PAGE and ligated to the 5' RNA adaptor that preferentially selects for 5' monophosphate substrates used in the small RNA cloning procedure. The ligation product and non-ligated small RNAs were resolved on 11% Urea-PAGE and subjected to northern blotting analysis. The ratio of small RNAs detected before and after linker ligation was quantified by ImageJ. 26G RNAs show similar levels of depletion after ligation compared to microRNAs miR-1 and miR-35, which are known to possess a 5' monophosphate. For miR-1, miR-35, and 26G-O1, a higher band corresponding to ligation product can be detected. The ~21nt endo-siRNAs appear to be poor substrates for the linker ligation.

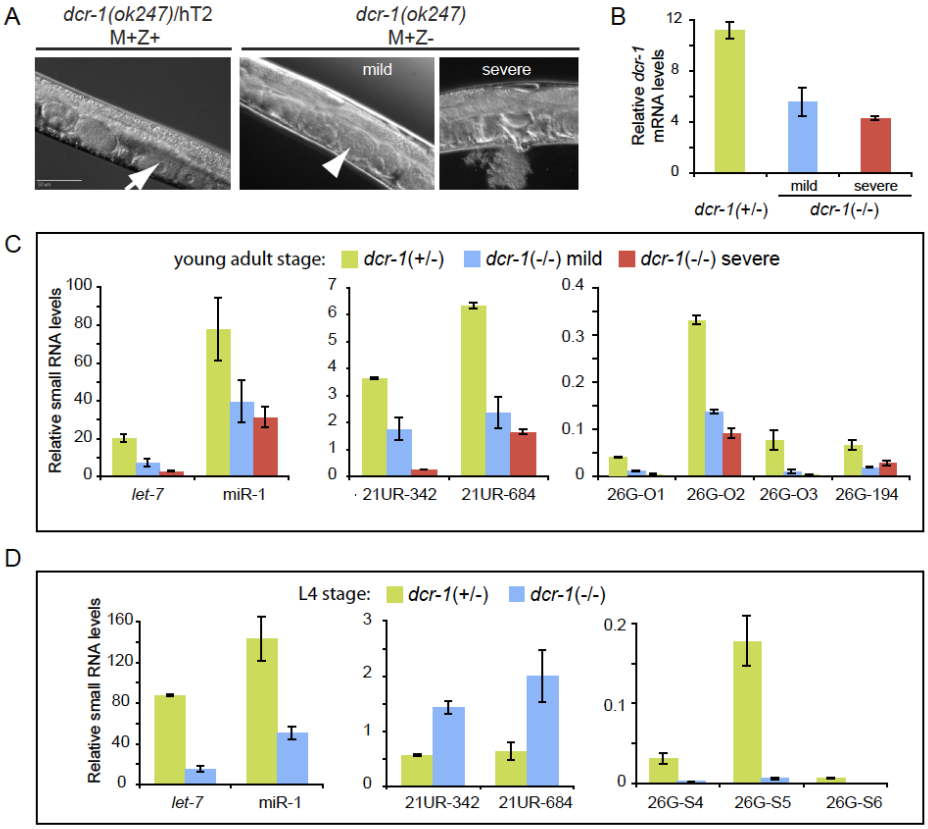


Figure 2.12 Expression of 26G RNAs are likely *dcr-1*-dependent.
A.) The population of *dcr-1(ok247)* null worms derived from heterozygous parents can be segregated into two groups: a "mild" phenotype group displaying abnormal oocytes (~100% penetrance) and a "severe" phenotype group displaying both abnormal oocytes and a bursting phenotype (<1% penetrance). An arrow indicates a normal oocyte in *dcr-1* (+/-) and an arrowhead indicates an abnormal oocyte in *dcr-1* (-/-). The *dcr-1* (+/-) heterozygotes retain both maternal and zygotic *dcr-1* mRNAs (M+Z+), while the *dcr-1* (-/-) homozygotes possess only maternal *dcr-1* mRNA inherited from the heterozygous parent (M+Z-). **B.)** *dcr-1* mRNAs are still present in *dcr-1*(-/-) animals due to maternal inheritance and slightly lower in the nulls displaying a "severe" phenotype versus the "mild" phenotype. The *dcr-1* mRNA levels were quantified by RT-qPCR and normalized to *act-1*. **C.)** At the young-adult stage, microRNAs (*let-7* and miR-1), 21U RNAs (21UR-342, 684), and class II oocyte/embryo 26G RNAs (26G-O1, O2, O3 and 194) are all depleted in *dcr-1* (-/-) relative to *dcr-1* (+/-). Relative levels of small RNAs were quantified by Taqman RT-qPCR and normalized to *act-1*. **D.)** At L4 larval stage, microRNAs (*let-7* and miR-1) and class I sperm 26G RNAs (26G-S4, S5, S6) are depleted in *dcr-1* (-/-) relative to *dcr-1* (+/-) while levels of 21U RNAs (21UR-342, 684) are slighted elevated in *dcr-1* (-/-).

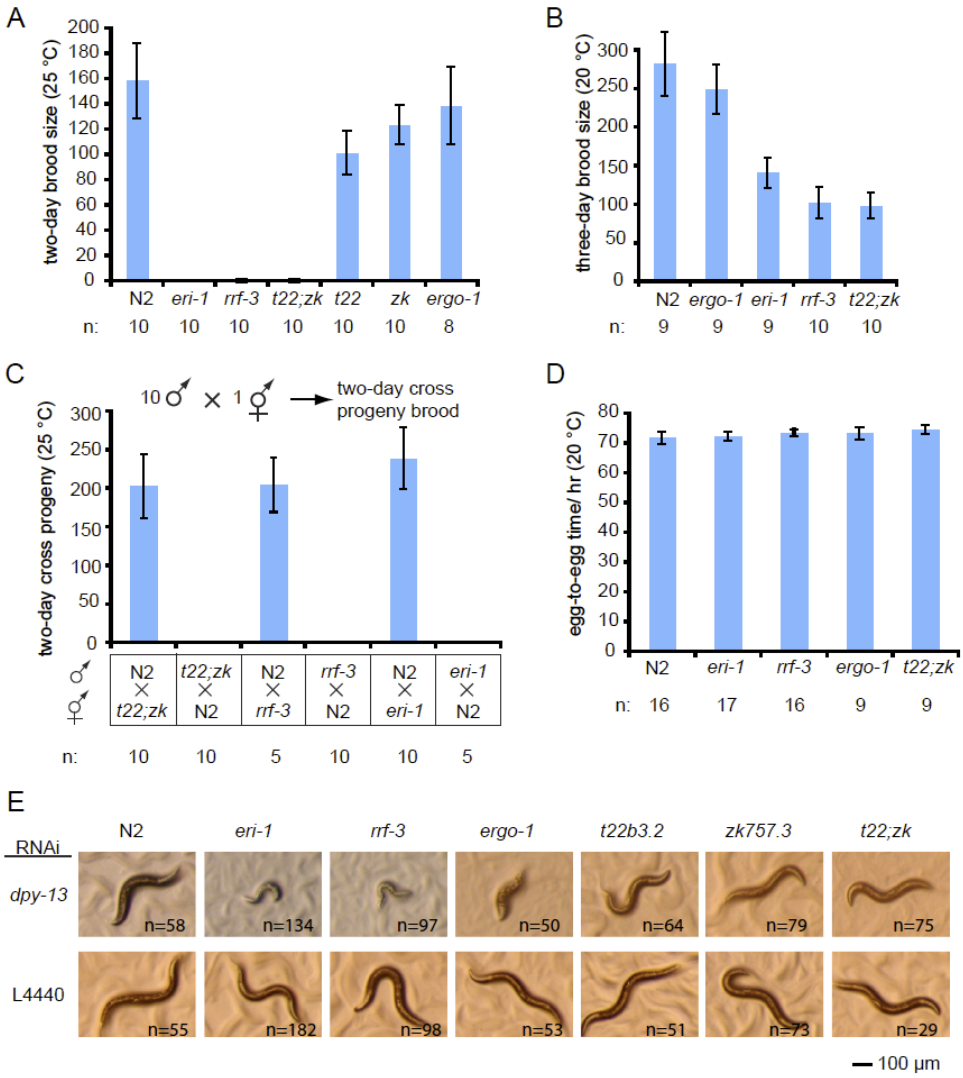


Figure 2.13 Phenotypes of mutants defective in 26G RNAs.
(A-B) The *t22b3.2(tm1155); zk757.3(tm1184)* double mutant is sterile at 25°C and exhibits significant loss of fertility at 20°C. Synchronized worms were singled at L4 stage and progeny brood size was counted for the subsequent two days. N is the number of parents assayed. Error bars represent standard deviation. Alleles used in this assay: *eri-1(mg366), rrf-3(pk1426)*, *ergo-1(tm1860)*, *t22b3.2(tm1155)*, *zk757.3(tm1184)*. **C.)** The *ts* sterility of *t22b3.2*; *zk757.3*, *eri-1*, and *rrf-3* can be fully rescued by crossing to WT males. For each cross, 10 males were crossed with 1 hermaphrodite, and two day cross progeny brood was scored. **D.)** N2, *t22b3.2(tm1155)*; *zk757.3(tm1184)*, *eri-1*, and *rrf-3* have similar egg-to-egg time at 20°C. The egg-to-egg time is the period it takes for fertilized eggs to develop into reproductive adults and produce fertilized eggs of their own. **E.)** The *t22b3.2(tm1155); zk757.3(tm1184)* double mutant does not display an enhanced RNAi phenotype. Synchronized L1 worms of indicated genotypes were subjected to feeding RNAi of *dpy-13* or control vector. L4 and young adult worms were examined for the severity of dumpy phenotype. A moderate dumpy phenotype was observed in N2, *t22b3.2(tm1155),*

zk757.3(tm1184), and the *t22b3.2(tm1155); zk757.3(tm1184)* double mutant. In contrast, RNAi inactivation of *dpy-13* in *eri-1(mg366), rrf-3(pk1426)*, and *ergo-1(tm1860)* generated a severe dumpy phenotype, indicating hypersensitivity to exogenous RNAi of *dpy-13*.

Chapter 3

Casein kinase 2 facilitates miRISC target binding and silencing in the C. elegans microRNA pathway2

3.1 Abstract

MicroRNAs (miRNAs) regulate diverse biological processes via repression of target mRNAs that contain complementary sites in their 3' untranslated regions (3' UTRs). While studies over the past two decades have provided fundamental insights into miRNA biogenesis and function, mechanisms controlling the activity of the miRNA-induced silencing complex (miRISC) are not well understood. Here we report the identification of a conserved serine/threonine kinase, casein kinase 2 (CK2) that functions in the miRNA pathway in *Caenorhabditis elegans*. Inactivation of CK2 results in developmental timing defects that phenocopy loss of the *let-7* family of miRNAs. In addition, CK2 broadly regulates the activities of multiple miRNAs in diverse developmental contexts, including *lsy-6* in neuronal fate choice, miR-84 in vulval precursor cell specification, and the miR-35 family of miRNAs in embryogenesis. By direct analysis of target mRNAs and proteins, we found that CK2 is required for the silencing of miRNA targets, including *lin-41*, *daf-12,* and *lin-14*. CK2 is dispensable for the accumulation of mature miRNAs

[2] Submitted to *PNAS* (under revision) with authors listed as Ting Han*, Vishal Khivansara*, James J. Moresco, Patricia G. Tu, John Yates III, and John K. Kim (* denotes equal contribution).

and the stability of core miRISC components. Instead, it is required for efficient association of target mRNAs with miRISC. Our study therefore reveals the novel role of CK2 in miRISC regulation at the level of target binding and silencing.

3.2 Introduction

Since the discovery of the first two miRNAs, *lin-4* and *let-7*, in *C. elegans*, members of the miRNA superfamily have emerged as conserved, ubiquitous regulators of gene expression critical for animal development, cell differentiation, apoptosis, and metabolism (15, 16, 18). To date, over 140 miRNAs have been identified in *C. elegans,* and ten times as many in human (http://www.mirbase.org). miRNAs are transcribed as long primary transcripts that are processed sequentially by RNase III enzymes Drosha and Dicer to produce a ~70 nt precursor hairpin and a ~22nt RNA duplex, respectively (8, 17). One strand of the duplex (the mature miRNA) is selectively loaded into an Argonaute family protein that forms the core of miRISC (8, 17).

Through partial base pairing with complementary sites predominantly located in the 3' UTRs of mRNAs, miRNAs guide miRISC to target mRNAs for degradation and/or translational inhibition (18). A unifying mechanism of target silencing remains to be resolved. Multiple studies suggest that target silencing occurs at various steps in translation, deadenylation, and mRNA decay (20). Two recent studies monitoring the temporal effects of miRNA-mediated regulation show that all these mechanisms contribute to silencing: miRNAs initially repress

translation by reducing the rate of initiation, followed by deadenylation and mRNA decay (24, 25).

Through a combination of genetic and biochemical approaches, several evolutionarily conserved miRISC components have been identified in *C. elegans*. *alg-1* and *alg-2* (Argonaute-like gene), orthologs of human Ago2, encode the *C. elegans* Argonautes that bind miRNAs (94). Mutation of *alg-1* results in pleiotropic phenotypes, implicating it in a broad range of biological pathways (94, 127). In contrast, *alg-2* mutants are superficially wild-type, suggesting a non-essential role in miRNA-mediated processes (127). ALG-1 physically interacts with AIN-1 and AIN-2 (ALG-1 interacting protein), functionally redundant orthologs of the human GW182 protein (128, 129). AIN-1/2 are required for the localization of ALG-1 to P-bodies, the major cytoplasmic centers for mRNA catabolism and storage, and function as a molecular link between target-bound miRISC and the downstream machinery for mRNA translational repression and degradation in P-bodies (128, 129). The conserved DEAD-box helicase CGH-1 is another key component of miRISC. Somatic CGH-1 physically interacts with ALG-1 and NHL-2, a TRIM-NHL ubiquitin ligase. Both *cgh-1* and *nhl-2* are required for miRISC activity, as mutation of either results in miRNA target desilencing (130). *C. elegans* miRISC also contains VIG-1 and TSN-1, homologs of factors identified in *D. melanogaster* RISC that are required for silencing of a *let-7* miRNA target reporter (131). In addition, a recent genome-wide RNAi screen yielded many more candidate miRNA pathway genes, suggesting that more factors required for miRISC function still await discovery (132).

The potential role of post-translational modifications in regulating miRISC activity remains to be fully explored. In this study, we identify a conserved serine/threonine kinase, casein kinase 2 (CK2), as a potent regulator of miRISC activity in *C. elegans*. We provide extensive genetic and molecular evidence that places CK2 downstream of miRNA biogenesis, at the step of target recruitment to miRISC.

3.3 Results

3.3.1 CK2 regulates developmental timing through the miRNA pathway.

We first identified *kin-10* in a genome-wide RNAi screen to discover novel factors in small RNA-mediated gene silencing pathways (133). *kin-10* encodes a subunit of the conserved serine/threonine kinase casein kinase 2 (CK2), which has roles in diverse biological pathways including signal transduction, transcriptional regulation, cell proliferation and differentiation, and tumorigenesis (134-136). The CK2 holoenzyme is composed of two catalytic α subunits and two regulatory β subunits in an α2β2 configuration (137). In *C. elegans*, *kin-3* and *kin-10* encode the catalytic and regulatory subunits of CK2, respectively (138). CK2 is ubiquitously expressed throughout *C. elegans* development (Figure 3.5), positioning it to play a universal role in regulating miRNA-mediated silencing.

In a complementary effort to identify novel factors in miRISC function, we immunopurified AIN-1 miRISC from adult stage wild-type and *ain-1*(*tm3681*) mutant animals and analyzed miRISC composition by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS). In addition to several core

miRISC factors (ALG-1, ALG-2, AIN-1, and CGH-1), we identified one peptide corresponding to KIN-3 (K.VLGTDELYEYIAR.Y; 3.6% sequence coverage) and two peptides corresponding to KIN-10 (R.GNEFFCEVDEEYIQDR.F and R.FNLTGLNEQVPK.Y; 12% sequence coverage) that were absent in proteins immunopurified from the *ain-1*(*tm3681*) mutant. However, we were unable to detect CK2 in the AIN-1 IP by immunoblotting, suggesting that CK2 might interact transiently with miRISC.

miRNAs control multiple aspects of developmental timing (i.e., the heterochronic pathway) in *C. elegans* (15, 16, 139-141). Members of the *let-7* family of miRNAs (*let-7*, miR-48, miR-84, and miR-241) share the same seed sequence, corresponding to positions 2-8 of the mature miRNAs, and therefore potentially regulate overlapping targets. Loss of the *let-7* family miRNAs or core miRISC factors leads to retarded heterochronic phenotypes including poorly formed alae, reiteration of larval stage seam cell divisions, failure to activate expression of adult-specific cuticle collagen COL-19, supernumerary molting cycles, and lethality (16, 94, 128, 129, 139-142). Because CK2 is essential for viability, we used RNAi to attenuate its expression (Figure 3.6) and examined the animals for defects in developmental timing.

Alae are cuticle protrusions that are only present in the L1 and dauer larval stages and the adult stage. The formation of adult alae was compromised upon *alg-1* RNAi (86% of animals), consistent with the role of ALG-1 in miRISC. 62-69% of CK2 RNAi animals exhibited malformed adult alae, with frequent incidence of severe disruption of straight alae ridges (Figure 3.1A). Similarly,

seam cell development was compromised upon CK2 RNAi. Lateral epithelial seam cells go through stereotypic symmetric or asymmetric divisions during larval transitions to generate 16 seam cells along the left and right sides of the adult animal that can be visualized by a GFP reporter expressed in seam cells (the P*scm*::*gfp* reporter) (143). Unlike wild-types adult animals that invariably have 16 seam cells, animals subjected to RNAi of CK2, *nhl-2*, or *ain-1* exhibited a supernumerary seam cell phenotype, with approximately 17 seam cells on average. *alg-1* RNAi animals displayed a more severe phenotype characterized by an average of 18 seam cells (Figure 3.1B). These results suggest that CK2 is required for the timing of seam cell divisions. COL-19 is an adult-specific collagen whose expression is controlled by the heterochronic pathway (144) and can be visualized by a P*col-19*:*gfp* reporter (142). Only 5% of adult animals under control RNAi displayed reduced P*col-19*::*gfp* expression in hypodermal cells. In contrast, 20-35% of CK2 RNAi animals showed reduced P*col-19*::*gfp* expression, comparable to *nhl-2* and *ain-1* RNAi. *alg-1* RNAi displayed a more penetrant phenotype (35-65% of animals showed reduced P*col-19*::*gfp* expression). Taken together, these results suggest that CK2 is required for the hypodermal remodeling during the larval to adult transition (Figure 3.1C).

Mild phenotypes resulting from weak mutations in the miRNA pathway genes (i.e., sensitized genetic backgrounds) can be enhanced by mutations in other genes required for the pathway (130). To examine further the role of CK2 in the miRNA pathway, we tested genetic interactions between CK2 and the *let-7* family of miRNAs and core miRISC genes. Between larval transitions, *C. elegans*

undergoes molting, which is preceded by a quiescent state, called lethargus. During lethargus, animals cease pharyngeal pumping and reduce locomotion. In retarded heterochronic mutants, lethargus is reiterated at the adult stage (16, 140, 141). The *mir-48* mutant exhibits a low-penetrant adult lethargus phenotype that can be exacerbated by deleting the miR-48 paralog, miR-84 (140). Using *mir-48*(*n4097*) as a sensitized genetic background, we found that the adult lethargus phenotype was dramatically enhanced by depletion of CK2 or core miRISC components (*nhl-2*, *alg-1*, or *ain-1*), while CK2 RNAi alone in the wild-type background was not sufficient to trigger adult lethargus (Figure 3.1D).

Loss of *let-7* activity at the larval to adult transition leads to a "bursting" phenotype, when the contents of the worm herniate through the vulva due to comprised vulval development (16). Two hypomorphic *let-7* alleles (*mg279* and *n2853*) exhibit a mild bursting phenotype that was enhanced by CK2 RNAi (Figure 3.1E). To rule out the possibility of off-target effects of RNAi, we further assayed *let-7* enhancement in CK2 deletion mutants. CK2 is essential for viability, yet the phenotypes associated with mutations in the two subunits varied in severity: *kin-3* (-/-) animals segregated from the *kin-3* (+/-) parent arrested at L3 stage, while *kin-10* (-/-) mutants arrested at L4 stage. As bursting can only be assessed at L4 or adult stages, we used *kin-10* (-/-) for this analysis. For two deletion mutants of *kin-10* (*ok1751* and *ok2031*), we observed enhanced bursting in *kin-10* (-/-) relative to *kin-10* (+/-), consistent with the *kin-10* RNAi results (Figure 3.1F). Three *let-7* family miRNAs, miR-48, -84, and -241, act redundantly to regulate the L2 to L3 transition; single, double, and triple mutants display

progressively more penetrant heterochronic phenotypes (140). CK2 RNAi enhanced the bursting phenotypes of both the single *mir-48*(*n4097*) mutant and the *mir-48*(*n4097*); *mir-84*(*n4037*) double mutant (Figure 3.1E). *alg-1*(*tm369*) and *ain-1*(*ku322*) mutants show only a mild bursting phenotype because of genetic redundancy with *alg-2* and *ain-2*, respectively (127, 129). CK2 RNAi also greatly exacerbated the bursting phenotype for both *alg-1* and *ain-1* mutants (Figure 3.1E). Consistent with the defect in silencing miRNA targets, the enhanced bursting phenotype in CK2 RNAi animals was suppressed by hypomorphic mutations in the *let-7* family targets, *lin-41* and *hbl-1* (Figure 3.1G, H). Taken together, these data indicate that CK2 genetically interacts with the *let-7* family of miRNAs to regulate developmental timing.

3.3.2 CK2 is required for the function of *lsy-6*, miR-35 family and miR-84 miRNAs.

Because CK2 is ubiquitously expressed (Figure 3.5), we tested whether it is broadly required for miRNA function. In the morphologically symmetric pair of ASE neurons (ASEL and ASER), the *lsy-6* miRNA functions to specify ASEL fate, which can be monitored by expression of an ASEL-specific P*lim-6*::*gfp* reporter (145). *lsy-6* null mutants invariably lack ASEL (i.e., no reporter GFP expression), but hypomorphic *lsy-6*(*ot150*) mutants display a partially penetrant ASEL fate specification defect. Because many neurons in *C. elegans* are refractory to RNAi (100), we performed RNAi experiments in the *nre-1* mutant background, which enhances neuronal RNAi (146). CK2 RNAi enhanced the penetrance of the

ASEL cell fate specification defect in *lsy-6*(*ot150*) animals to the same extent as *alg-1* RNAi (Figure 3.2A), indicating that CK2 is required for the function of the *lsy-6* miRNA.

The miR-35 family of miRNAs (miR-35 to -42) is expressed during embryogenesis and functions redundantly to control embryonic development (147). Deletion of seven out of the eight family members (miR-35 to -41) leads to a temperature-sensitive late embryonic or early L1 lethal phenotype (147). At 15°C, about 10% of embryos display this phenotype, while the remaining 90% develop normally. Inhibition of CK2, *nhl-2*, or *alg-1* dramatically increased embryonic or L1 lethality, consistent with the requirement of CK2 for the function of the miR-35 family of miRNAs (Figure 3.2B).

In addition to regulating developmental timing, miR-84, a member of the *let-7* family of miRNAs, also regulates vulval precursor cell (VPC) fate specification (148). In L3 larvae, all of the six multipotent VPCs (P3-8.p) have the capacity to adopt the 1° cell fate; however, an inductive signal from the gonadal anchor cell activates LET-60/RAS signaling only in P6.p to specify the 1° vulval cell fate (149). In P5.p and P7.p, LET-60 activity is further repressed by miR-84, and failure to repress LET-60 activity leads to a multivulva (Muv) phenotype (148). *let-60*(*ga89*) is a gain of function mutant with low-penetrant Muv phenotype (150). CK2 RNAi enhanced the *let-60*(*ga89*) Muv phenotype (Figure 3.2C), consistent with the model that CK2 promotes miR-84 function. Taken together, our data are consistent with a general role of CK2 in diverse biological processes mediated by miRNAs.

3.3.3 CK2 is dispensable for miRNA biogenesis or expression of core miRISC factors.

To investigate if CK2 is required for miRNA biogenesis, we performed northern blotting to examine miRNA levels in CK2 RNAi animals. We saw no difference in mature miRNA levels for miR-48, miR-1, *let-7*, and miR-58 (Figure 3.3A and Figure 3.7) in wild-type versus CK2 RNAi animals, or in the genetically sensitized *alg-1*(*tm369*) versus *alg-1(tm369);* CK2 RNAi animals (Figure 3.3A), indicating that CK2 functions downstream of miRNA biogenesis. In order to assess the impact of CK2 on the stability of miRISC components, we performed western blotting to examine levels of ALG-1, AIN-1, CGH-1, VIG-1, and TSN-1, and detected no difference in wild-type versus CK2 RNAi animals (Figure 3.3B), indicating that the stability of individual miRISC factors is unaffected by CK2 depletion.

3.3.4 CK2 is required for target silencing.

In *C. elegans*, the *let-7* family of miRNAs triggers degradation of its target mRNAs, *lin-41* and *daf-12* (19, 151). LIN-41 is a Ring finger-B Box-Coiled coil (RBCC) protein and a member of the NHL (NCL-1, HT2A, and LIN-41) family of proteins. *lin-41* mRNA is down-regulated by *let-7* at the L4 to adult transition (19, 152). *daf-12* encodes a nuclear steroid hormone receptor that integrates environmental signals and developmental timing and is down-regulated by the *let-7* family of miRNAs starting from the L2 to L3 transition to the L4 to adult

transition (153, 154). *let-7*(*n2853*) mutants and *alg-1* RNAi animals exhibited elevated *lin-41* and *daf-12* mRNA levels that were between 1.5- and 4-fold higher than wild-type (Figure 3.3C, D). Similarly, knockdown of CK2 increased *lin-41* and *daf-12* mRNA levels by 1.5- to 2-fold (Figure 3.3C, D). We also compared *kin-10* (-/-) versus *kin-10* (+/-) genetic mutants for defects in silencing *let-7* family target mRNAs and observed significant increase in expression of *lin-41* mRNA, consistent with the CK2 RNAi results (Figure 3.3C).

To determine if derepression of target mRNAs by CK2 RNAi also promotes elevated target mRNA translation, we examined the protein levels of LIN-14 upon CK2 RNAi. The miRNA *lin-4* is required for the L1 to L2 transition by silencing the expression of *lin-14*, which encodes a novel nuclear protein. *lin-14* is expressed at high levels in newly hatched L1 animals and is downregulated by *lin-4* at the L2 stage (15, 155). We analyzed endogenous LIN-14 protein levels at L2 stage by immunoblotting and observed a 2- to 3-fold increase in LIN-14 protein levels in CK2 RNAi animals relative to vector control (Figure 3.3E). Taken together, these data indicate that miRNA-mediated silencing of *lin-41* and *daf-12* mRNAs and regulation of LIN-14 protein levels require CK2.

3.3.5 CK2 promotes target association with miRISC

Target mRNA recruitment to miRISC is a critical step in miRNA-mediated silencing. Defects in miRISC activity could result from compromised ability to bind to miRNAs and/or their targets. To determine if CK2 is required for miRNA and target binding to miRISC, we performed RNA-immunoprecipitation (RIP)

assays in a strain expressing a rescuing *gfp*::*alg-1* transgene (156) in CK2 RNAi versus vector control or *gfp* RNAi animals (Figure 3.4A). We spiked in a firefly luciferase mRNA to RIP before RNA extraction to control for RNA extraction efficiency, and miRNA and target mRNA abundance in each RIP sample was normalized to luciferase mRNA. By RT-qPCR, we observed no significant changes in the amount of *let-7*, miR-48, or miR-1 in the RIP from CK2 RNAi animals compared to vector control (Figure 3.4B). However, *lin-41*, *daf-12*, and *mef-2* (a target of miR-1(157)) mRNAs showed 2-fold reduction in the RIP relative to control (Figure 3.4C). GFP RNAi, which significantly decreases the expression of GFP::ALG-1, also reduces both miRNA and target mRNA levels in the RIP (Figure 3.4B,C). Taken together, these data indicate that CK2 promotes efficient binding of target mRNAs to miRISC but is dispensable for miRNA binding to miRISC.

3.4 Discussion

Our finding that inactivation of CK2 phenocopies mutants of core miRISC factors, such as ALG-1, AIN-1, NHL-2, and CGH-1, suggests that CK2 regulates miRISC activities. As CK2 is dispensable for miRNA biogenesis or loading into miRISC, yet required for miRISC target association and silencing, we propose that CK2 may modulate the assembly or conformation of miRISC. Absence of CK2 may compromise the interactions among miRISC components and lead to weakened association with target mRNAs. The observation that loss of CK2 de-silences target mRNAs and increases target translation supports a model

whereby wild-type miRISC sequesters target mRNAs from active translation. Accordingly, depletion of CK2 would liberate mRNAs from miRISC control and lead to an increase in translational efficiency and target mRNA stability.

To date, a limited number of reports implicate post-translational modifications in the regulation of miRISC components (158). In particular, Argonautes appear to be major substrates for modification. Hydroxylation of human Ago2 at proline-700 is required for its stability, P-body localization, and small RNA-mediated silencing activity (159). Blocking phosphorylation of human Ago2 at serine-387 with a p38 MAPK inhibitor reduces Ago2 localization to P-bodies (160). Finally, ubiquitination of mouse Ago2 by the TRIM-NHL protein Lin41 promotes its degradation via the proteasome (161).

Most kinases in vivo are only active upon receiving correct stimuli such as ligand binding or phosphorylation by an upstream priming kinase. Despite considerable efforts, the matter of whether CK2 is constitutively active or activated by specific signals still remains controversial (162). As CK2 resides in many subcellular compartments and could phosphorylate many different substrates, recent studies have proposed that CK2 could be regulated through protein-protein interactions to phosphorylate specific substrates (163). We speculate that CK2 could be transiently recruited to miRISC and phosphorylate one or more miRISC factors to enhance miRISC activity. CK2 phosphorylates serines or threonines located within the consensus motif S/TXXD/E. We queried a published *C. elegans* phosphoproteome dataset (164) and found peptides phosphorylated at CK2 motifs in CGH-1 and VIG-1, suggesting that these

miRISC components may be CK2 substrates. The CK2-dependent phosphorylation of miRISC factors is currently under investigation. As modulating miRNA activity is crucial for development and disease, dissecting the mechanisms that activate or antagonize CK2 activity in the miRNA pathway may also provide further insights into how the miRNA pathway is regulated.

3.5 Materials and Methods

Nematode strains and culture methods. Worm strains were grown and maintained at 20°C as described (124), except where otherwise indicated. All strains used for this study are listed in Table 3.1. To synchronize worms, embryos isolated by standard alkaline/hypochlorite treatment (124) were hatched overnight to obtain arrested L1 larvae.

Plasmids and transgenic strains. The *kin-3*::*gfp* reporter construct (pJK194) was generated by introducing the following fragments into pJK211 (a vector derived from pPD49.26 from the Fire vector kit): a 2 kb fragment upstream of *kin-3*, a 2.2 kb PCR fragment containing *kin-3* genomic coding region (without termination codon), a *gfp* coding region (0.9 kb fragment with multiple synthetic introns and termination codon), and a 1.2 kb PCR fragment immediately downstream of the *kin-3* termination codon. DNA transformation and microinjection were performed as described (165).

RNAi. All RNAi clones were picked from the Ahringer RNAi library (166) except the *kin-3* RNAi clone, which was picked from the Vidal RNAi library (167). RNAi experiments were performed at 20°C (except the *lsy-6* assay (25°C), miR-

35 assay (15°C), and *let-60* assay (25°C)) starting from synchronized L1 larvae according to the standard feeding RNAi protocol (133). Since several genes in this study are essential for viability, we optimized the duration (1 or 2 generations on RNAi) and strength of RNAi (by dilution of the bacterial RNAi clones with bacteria harboring the empty vector *L4440*) to achieve efficient knockdown in viable animals. For RNAi spanning two generations, two rounds of synchronization were performed. RNAi conditions for alae, seam cell, P*col-19*::*gfp*, lethargus, bursting, and *let-60(ga89)* Muv assays were as follows: *kin-3* RNAi (1 generation; undiluted), *kin-10* RNAi (2 generations; 1/2 strength by diluting with vector culture), *alg-1* RNAi (1 generation; 1/2 strength), *nhl-2* RNAi (2 generations; undiluted), and *ain-1* RNAi (2 generations; undiluted). For the *lsy-6* assay, synchronized worms were first cultured on vector RNAi for 36 h and then transferred to full strength RNAi plates of the indicated genes for 120 h. Progeny were then scored for GFP expression in the ASEL neuron. For the miR-35 assay, 15-20 L3-stage worms were transferred to full strength RNAi plates and cultured for 60 h to reach adulthood. After egg laying, adult worms were picked off the plates, and the hatchlings were scored 36 h later.

Microscopy and phenotypic analyses. Olympus BX61 epifluorescence compound microscope with Nomarski optics was used to examine the alae defect, and a Leica MZ16 F fluorescence stereomicroscope was used to score seam cell numbers (P*scm*::*gfp* GFP reporter) and adult collagen expression (P*col-19*::*gfp* GFP reporter). To monitor adult lethargus, cessation of pharyngeal pumping was examined in synchronized populations of worms every 2 h from 60

h to 72 h post recovery from L1 diapause. At each time point, worms undergoing lethargus were removed from the RNAi plates to prevent double counting.

RNA analyses. RNAi conditions for empty vector, *kin-3*, *kin-10*, and *alg-1* are described above. miRNA levels were assayed from worms grown for 48 h post L1; *lin-41* mRNA levels were assayed from worms grown for 44 h post L1; *daf-12* mRNA levels were assayed from worms grown for 40 hr post L1. Total RNA isolation was carried out using TriReagent (Ambion) following the vendor's protocol with the following modifications. Samples in TriReagent were subjected to three rounds of freeze-thaw-vortex cycles to improve RNA extraction efficiency. RNAs were precipitated in isopropanol for 1 h at -80°C followed by three washes with 70% ethanol. Northern blotting was performed as described (168) using 1.5 µg total RNA and Starfire DNA probes (IDT) (Table 3.2). For quantification of mRNAs, 250 ng of total RNAs were converted into cDNAs using Multiscribe Reverse Transcriptase (Applied Biosystems) following the vendor's protocol. qPCR was performed using Power SYBR Green master mix 2X (Applied Biosystems). Primers used for this study are listed in Table 3.2. *eft-2* mRNA levels were used for normalization. For RT-qPCR analysis of miRNAs, miRNA Taqman assays (Applied Biosystems) were performed following the vendor's protocol.

Protein analyses. Custom rabbit polyclonal antibodies were generated by Proteintech Group, Inc. using peptide antigens (listed in Table 3.2) conjugated to Keyhole limpet hemocyanin (KLH) carrier protein. LIN-14 antibody was a gift from Gary Ruvkun (169). For western blotting, proteins were isolated from

synchronized late L4 worms by direct boiling in Novex Tris-Glycine SDS sample buffer (Invitrogen). Protein samples were resolved by Novex Tris-Glycine gel (Invitrogen), transferred onto Immobilon-FL transfer membrane (Millipore), and probed with rabbit polyclonal antibodies to ALG-1 (1:1000), AIN-1 (1:1000), CGH-1 (1:1000), TSN-1 (1:1000), VIG-1 (1:1000), or γ-tubulin (Sigma LL-17) (1:2000). Peroxidase-AffiniPure goat anti-rabbit IgG secondary antibody was used at 1:10,000 (Jackson ImmunoResearch Laboratories) for detection using Pierce ECL Western Blotting Substrate (Thermo Scientific). Western blot analysis to determine LIN-14 protein expression was performed as described (170).

Immunoprecipitation of AIN-1 complex and mass spectrometry. Wild-type and *ain-1*(*tm3681*) adult worms were frozen in liquid nitrogen and homogenized with a Mixer Mill MM 400 ball mill homogenizer (Retsch). Homogenates were suspended in lysis buffer (50 mM HEPES pH 7.4, 1 mM EGTA, 1 mM $MgCl_2$, 100 mM KCl, 10% glycerol, 0.05% NP-40 supplemented with Complete, Mini, EDTA-free Protease Inhibitor Cocktail tablet (Roche Applied Sciences)) and clarified by centrifugation at 12,000X g for 12 min at 4°C. For immunoprecipitations, adult homogenates were incubated at 4°C for 4 h with 75 µg anti-AIN-1 rabbit polyclonal antibody conjugated to Dynabeads Protein A (Invitrogen), after which the beads were washed three times with wash buffer (50 mM HEPES pH 7.4, 1 mM EGTA, 1 mM $MgCl_2$, 300 mM KCl, 10% glycerol, 0.05% NP-40 supplemented with Complete, Mini, EDTA-free Protease Inhibitor Cocktail tablet (Roche Applied Sciences)). The immunoprecipitated proteins were eluted from the beads with three aliquots of 150 µL of 0.1 M glycine, pH 2.6 (450

µL of total eluates). Eluates were neutralized with 150 µL of 2M Tris-HCl, pH 8.5, combined with 1/5 volume of 100% trichloroacetic acid and precipitated overnight at 4°C. Proteins were pelleted by centrifugation at 20,000X g for 30 min and washed twice with acetone before mass spectrometry analysis. Protein identification by mass spectrometry was performed as previously described (171).

RNA-Immunoprecipitation (RIP). GFP-ALG-1 was immunoprecipitated from synchronized L4 stage CT20 (*gfp*::*alg-1*) worms using a mouse monoclonal anti-GFP (3E6) (Invitrogen). For each RIP, 10 µg of anti-GFP antibody was cross-linked to Dynabeads Protein A (Invitrogen) and incubated with lysate prepared from 0.3 ml of frozen worms at 4°C for 1 h. Beads were washed four times with RIP wash buffer (50 mM Tris-HCl pH 7.5, 200 mM KCl, 0.05% NP-40) and then split into two aliquots for protein and RNA analyses. For protein analyses, 30 µL of 1X Tris-glycine SDS sample buffer (Invitrogen) was added directly to the beads and incubated at 50°C for 10 min. The eluted proteins were transferred to a new tube, supplemented with 0.1 M DTT, and incubated at 90°C for 5 min before western blotting. For RNA analysis, 1 ml of TRI-Reagent (Ambion) and 10 ng of Firefly Luciferase RNA (Promega) was directly added to the beads and incubated at room temperature for 5 min. RNAs were precipitated in isopropanol for 2 h at -30°C followed by three washes with 70% ethanol. For RT-qPCR quantification of miRNAs and target mRNAs in RIP, *eft-2* and firefly luciferase mRNA levels (Table 3.2) were used for normalization of crude lysate and IP samples, respectively.

3.6 Acknowledgements

We thank Allison Billi, Mallory Freeberg, and Amelia Alessi for helpful comments on the manuscript. We thank the *Caenorhabditis* Genetics Center, Shohei Mitani, and Frank Slack for strains; Gary Ruvkun for strains and the LIN-14 antibody. This research was supported by National Institute of General Medical Sciences (NIGMS) R01GM088565 and the Pew Charitable Trusts (J.K.K.); Rackham Pre-doctoral fellowship from University of Michigan (T.H.); and the National Center for Research Resources 5P41RR011823-17, NIGMS 8 P41 GM103533-17, and the National Institute on Aging R01AG027463-04 (J.R.Y).

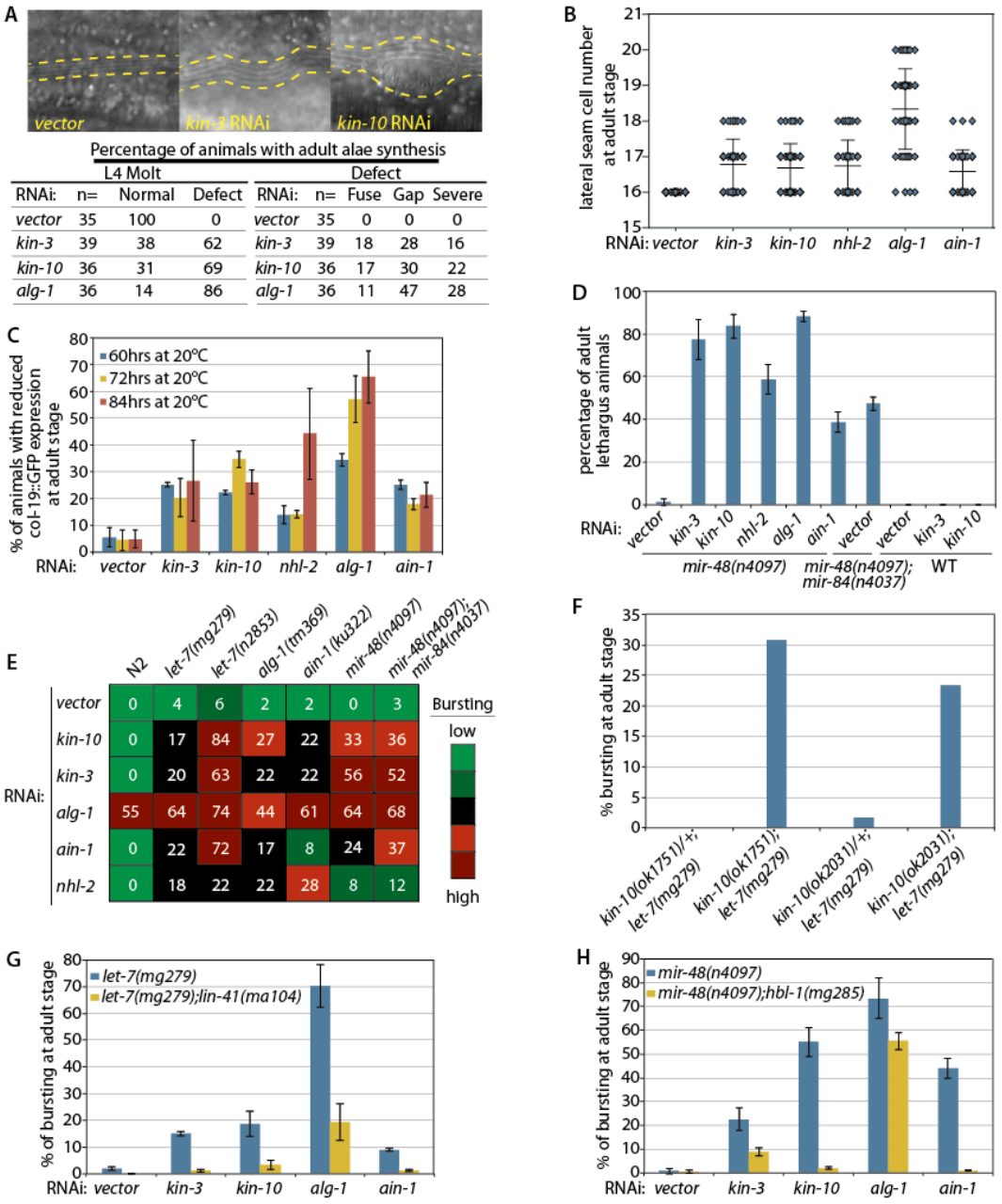


Figure 3.1 Inactivation of CK2 results in retarded heterochronic phenotypes associated with the *let-7* family of miRNAs.

(A) CK2 RNAi exhibits malformed adult alae. Representative Nomarski DIC images of adult alae (60 h post L1 at 20°C) are shown. Statistics of alae defects are listed in the table. (B) CK2 RNAi compromises the timing of seam cell divisions (n=50 for each RNAi treatment; RNAi of *kin-3*, *kin-10*, *nhl-2*, *alg-1*, and *ain-1* all exhibit significantly different outcomes from vector control, p<0.001 by two tailed Student's t-test). (C) CK2 RNAi reduces P*col-19*::*gfp* reporter expression. GFP reporter expression was scored at 60 h, 72 h, 84 h post L1 at 20°C. Mean and standard deviation (SD) were plotted from two biological

replicates (n>113 for each RNAi treatment). (D) CK2 RNAi enhances the adult lethargus phenotype of *mir-48*(*n4097*). Percentage of adult animals inappropriately entering lethargus were scored from 60 h to 72 h post L1 at 20°C. Mean and standard deviation (SD) were plotted from two biological replicates (n>89 for each RNAi treatment). (E) CK2 RNAi enhances the bursting phenotype of several miRNA pathway components. Percentages of bursting were visualized as a heat map. Bursting was scored at 72 h post L1 at 20°C (two biological replicates were performed; n>102 for each replicate). (F) Enhancement of *let-7* bursting in *kin-10* (-/-) segregated from *kin-10* (+/-) parent at 20°C. (G) *lin-41(ma104)* suppresses bursting in *let-7*(*mg279*); CK2 RNAi. (H) *hbl-1(mg285)* suppresses bursting in *mir-48*(*n4097*); CK2 RNAi.

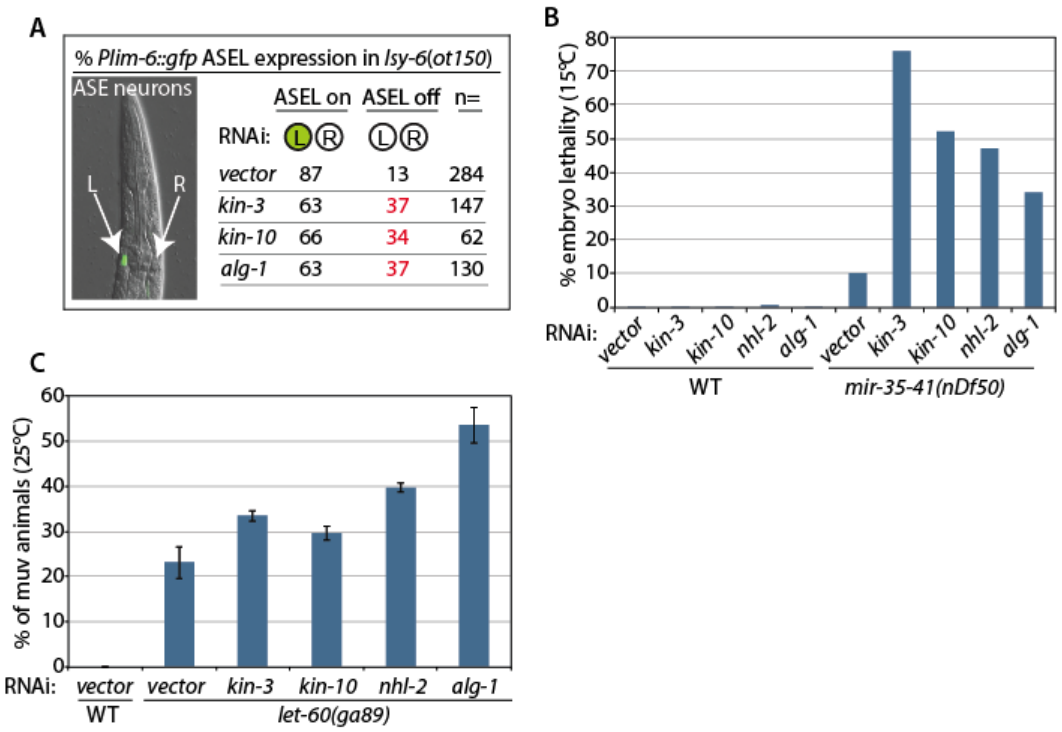


Figure 3.2 CK2 is required for the activities of *lsy-6*, miR-35, and miR-84 family miRNAs.
(A) *lsy-6* miRNA is required for ASEL neuron specification in wild type animals indicated by the expression of P*lim-6*::*gfp*. CK2 RNAi increases the penetrance of ASEL specification defect in the hypomorphic *lsy-6*(*ot150*) mutant. (B) CK2 RNAi enhances the embryonic lethality associated with *mir-35-41*(*nDf50*). (C) CK2 RNAi enhances the Muv phenotype associated with the gain-of-function *let-60*(*ga89*) mutant. Mean and SD were plotted for two biological replicates (n>94 for each replicate; for *let-60*(*ga89*), RNAi of *kin-3*, *kin-10*, *nhl-2*, *alg-1* all exhibit significantly different outcomes from vector control, $p<0.001$ by χ^2 test).

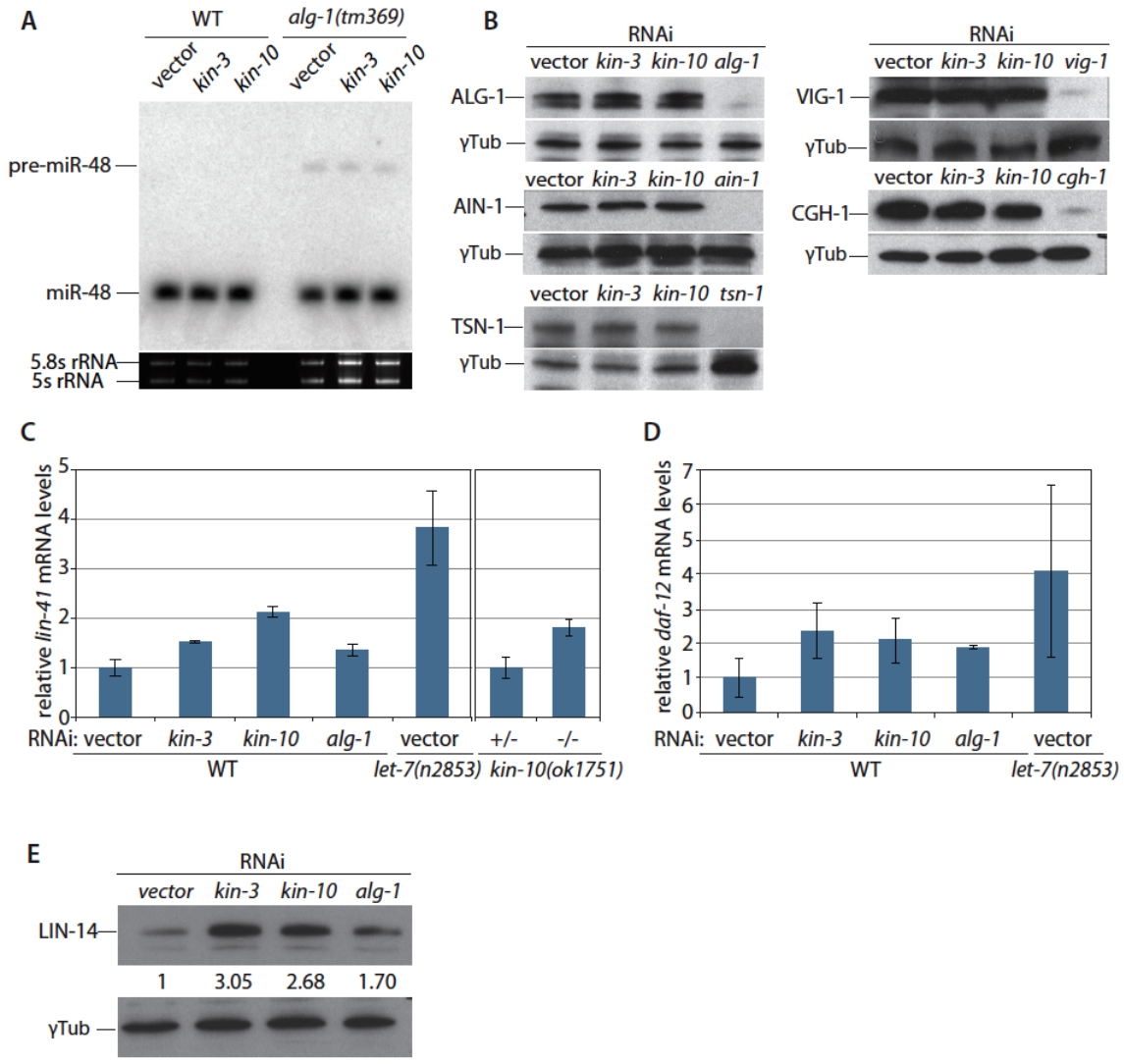


Figure 3.3 CK2 is dispensable for miRNA biogenesis and miRISC factor expression but is required for target silencing.
(A) miRNA northern blotting in wild-type and sensitized *alg-1*(*tm369*) backgrounds shows that CK2 RNAi does not alter the expression of miR-48 precursor or mature miRNA levels. (B) CK2 RNAi does not affect the protein levels of core miRISC factors. (C) CK2 RNAi elevates *lin-41* mRNA levels at L4 stage (44 h post L1, 20°C), and *kin-10* (-/-) mutant exhibits increased *lin-41* mRNA levels relative to *kin-10* (+/-) at L4 stage (44 h post L1, 20°C). Mean and 95% confidence interval (CI) were plotted for two biological replicates. (D) CK2 RNAi elevates *daf-12* mRNA levels at late L3 stage (40 h post L1, 20°C). Mean and 95% confidence interval (CI) were plotted for two biological replicates. (E) Western blot of LIN-14 at L2 stage (20 h post L1, 20°C) after indicated RNAi treatment. γ-tubulin was used as a loading control. Quantification of band intensities was performed using ImageJ.

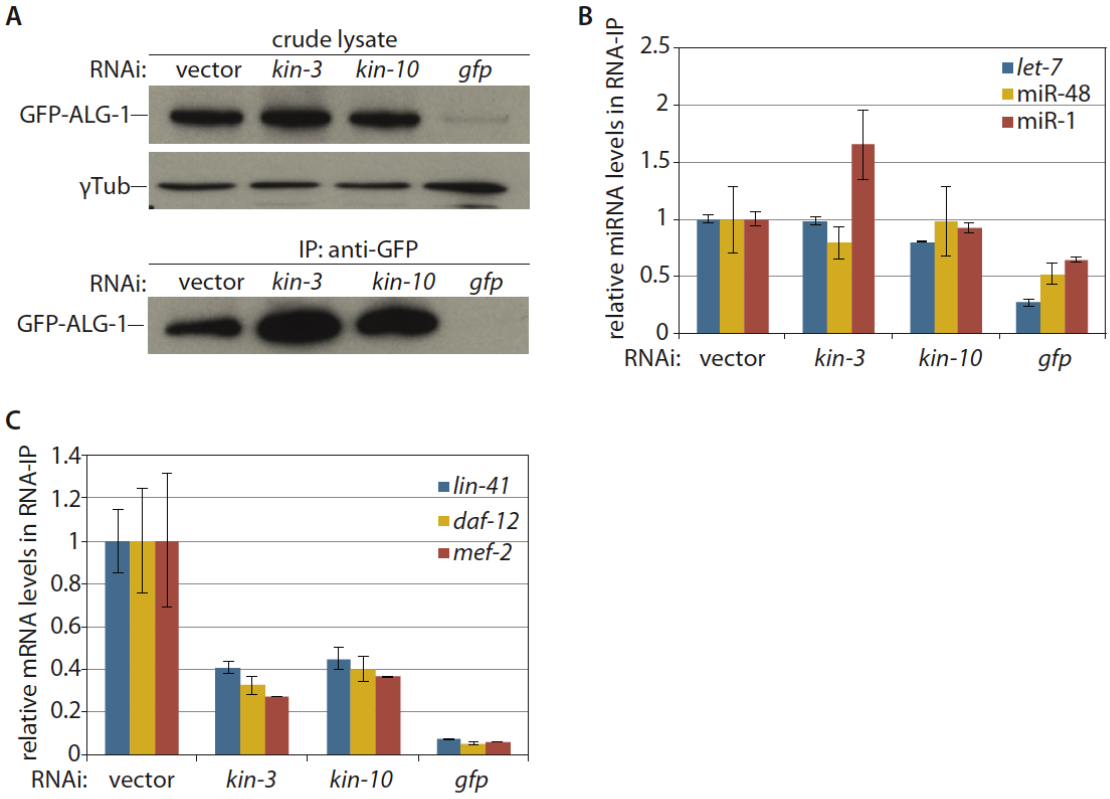


Figure 3.4 CK2 is required for recruitment of target mRNAs to miRISC.
(A) Detection of GFP::ALG-1 proteins by western blotting in crude lysates and IPs from L4 stage animals (48 h post L1, 20°C) following indicated RNAi treatment. γ-tubulin was used as a loading control. (B) *let-7*, miR-48, and miR-1 levels remained unchanged in RNA-IP in CK2 RNAi relative to wild-type. miRNA levels were normalized to a spiked-in firefly luciferase mRNA control. Mean and 95% confidence interval (CI) were plotted for two technical replicates. Similar results have been observed for five independent experiments. (C) CK2 RNAi reduces the amount of *lin-41*, *daf-12*, and *mef-2* mRNAs bound to GFP-ALG-1 compared to vector control. Mean and 95% confidence interval (CI) are shown for two technical replicates. Similar results have been observed for five independent experiments.

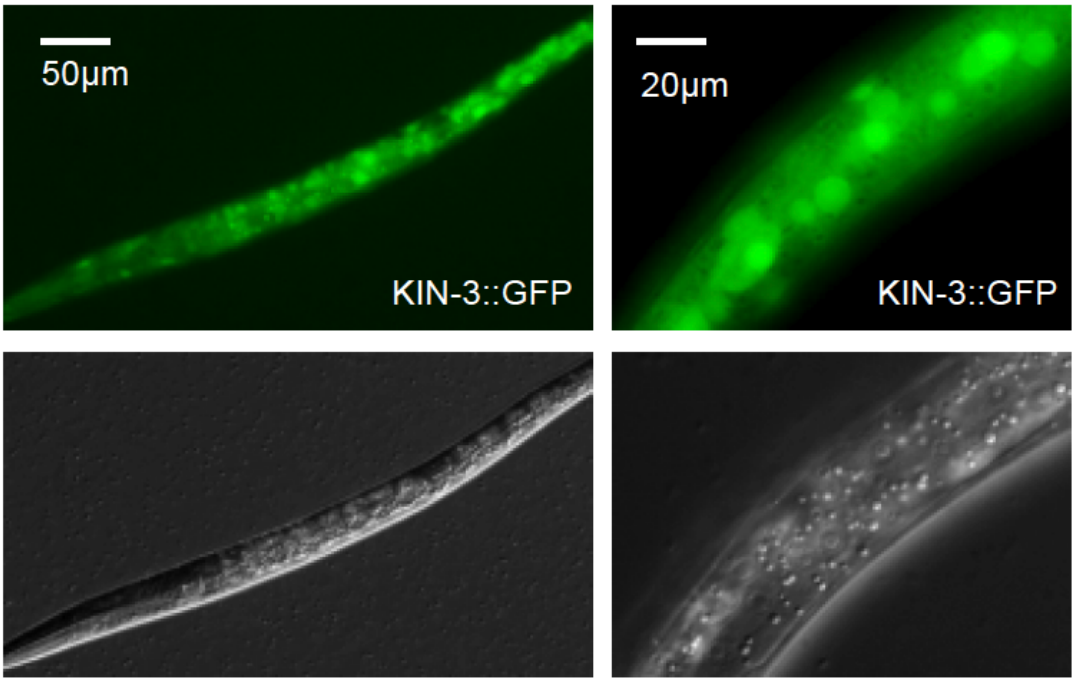


Figure 3.5 KIN-3 is ubiquitously expressed.
Fluorescence microscopy of KIN-3::GFP driven by a 2 kb *kin-3* endogenous promoter.

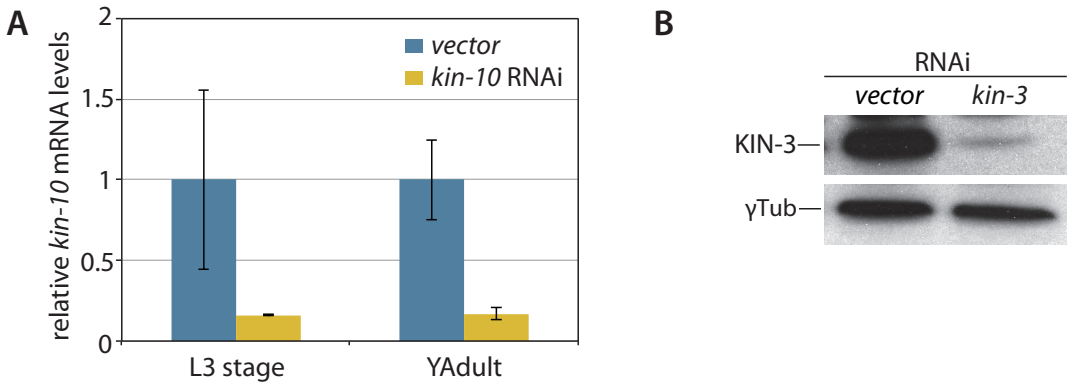


Figure 3.6 RNAi efficiently knocks down CK2 expression.
(A) *kin-10* mRNAs were efficiently depleted by *kin-10* RNAi. (B) KIN-3 proteins were depleted in *kin-3* RNAi relative to wild-type.

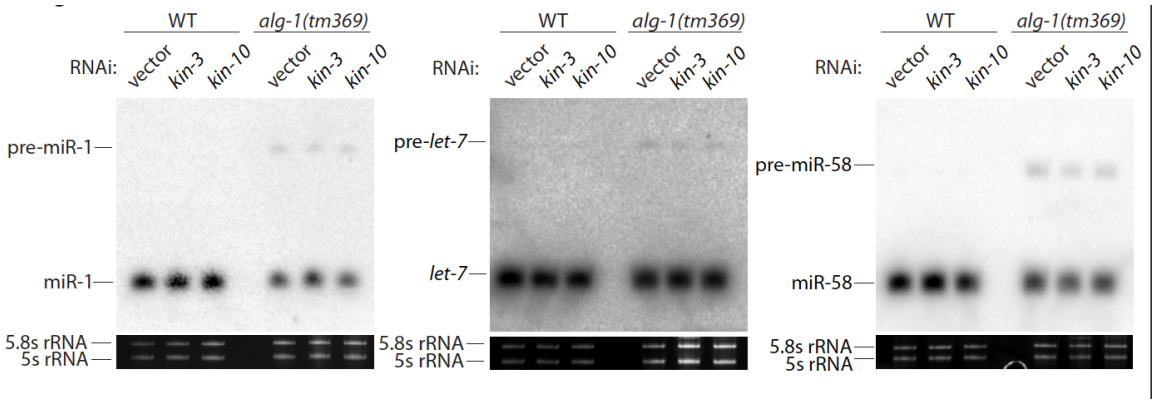


Figure 3.7 CK2 is not required for miRNA biogenesis.
Three more examples (miR-1, *let-7*, and miR-58) are shown here in addition to miR-48 shown in Figure 3.3A.

Table 3.1 Strains used in this study.

N2	*C. elegans* wild isolate (wild-type reference)	Source
MT7626	*let-7(n2853)*	CGC
GR1432	*let-7(mg279)*	CGC
MT13650	*mir-48(n4097)*	CGC
MT13652	*mir-48 (n4097); mir-84(n4037).*	CGC
CT8	*lin-41(ma104)*	CGC
MT13651	*mir-84(n4037)*	CGC
CT11	*hbl-1(mg285)*	CGC
QK003	*mir-48(n4097);hbl-1(mg285)*	this study
QK004	*let-7(mg279); lin-41(ma104)*	this study
tm369	*alg-1(tm369)*	Shohei Mitani
MH2385	*ain-1(ku322)*	CGC
JR672	*Is54[scm::gfp]*	Gary Ruvkun
SD551	*let-60(ga89)*	CGC
MT14119	*mir-35 to -41(nDf50)II*	CGC
OH3646	*otIs114[Plim-6::gfp] I; lsy-6(ot150) V*	CGC
VH624	*nre-1(hd20) lin-15b(hd126) X; rhIs13[unc-119::gfp, dpy-20(+)]*	CGC
QK005	*otIs114[Plim-6::gfp] I; lsy-6(ot150) V; nre-1(hd20) lin-15b(hd126) X*	this study
RG365	*him-1(e879) I; veIs13[col-19::gfp; rol-6(su1006)] V*	Gary Ruvkun (originally from Ann Rougvie)
CT20	*Is[alg-1::GFP-alg-1]*	Frank Slack
VC1280	*kin-10(ok1751) I/hT2[bli-4(e937) let-?(q782) qIs48](I;III).*	CGC
VC1609	*kin-10(ok2031) I/hT2[bli-4(e937) let-?(q782) qIs48](I;III).*	CGC
QK006	*Is[kin-3::kin-3::GFP]*	this study

Table 3.2 Oligo and peptide sequences used in this study.

DNA oligo sequences

ACGCTCGTGATGAGTTCAAG	*eft-2* qPCR forward
ATTTGGTCCAGTTCCGTCTG	*eft-2* qPCR reverse
GGTTCCAAATGCCACAAGAG	*lin-41*qPCR forward
AGGTCCAACTGCCAAATCAG	*lin-41*qPCR reverse
GATCCTCCGATGAACGAAAA	*daf-12* qPCR forward
CTCTTCGGCTTCACCAGAAC	*daf-12* qPCR reverse
CTCACTGAGACTACATCAGC	firefly luciferase qPCR forward
TCCAGATCCACAACCTTCGC	firefly luciferase qPCR reverse

Starfire Probe sequences

TACATACTTCTTTACATTCCA	miR-1
AACTATACAACCTACTACCTCA	*let-7*
TCGCATCTACTGAGCCTACCTCA	miR-48
TGCCGTACTGAACGATCTCA	miR-58

Antigenic peptide sequences

QAGSLAPGVPIGNTSVSI(C)	ALG-1
WGDPPLSDVQYPLQPHASF(C)	AIN-1
(C)IEPIPKTVDPKLYVADQQLVDA	CGH-1
(C)GRNNTPFNASDDAFPALGAK	VIG-1
(C)AEGLALADHRREPRLQTLVNDY	TSN-1
MPPIPSRARVYAEVNPSRP(C)	KIN-3

Chapter 4

The landscape of *C. elegans* 3'UTRs[3]

4.1 Abstract

Three-prime untranslated regions (3′UTRs) of metazoan messenger RNAs (mRNAs) contain numerous regulatory elements, yet remain largely uncharacterized. Using polyA capture, 3′ rapid amplification of complementary DNA (cDNA) ends, full-length cDNAs, and RNA-seq, we defined ~26,000 distinct 3′UTRs in *Caenorhabditis elegans* for ~85% of the 18,328 experimentally supported protein-coding genes and revised ~40% of gene models. Alternative 3′UTR isoforms are frequent, often differentially expressed during development. Average 3′UTR length decreases with animal age. Surprisingly, no polyadenylation signal (PAS) was detected for 13% of polyadenylation sites, predominantly among shorter alternative isoforms. Trans-spliced (versus non–trans-spliced) mRNAs possess longer 3′UTRs and frequently contain no PAS or variant PAS. We identified conserved 3′UTR motifs, isoform-specific predicted

[3] Originally published in *Science* (2010;329(5990):432-5) with authors listed as Marco Mangone*, Arun Prasad Manoharan*, Danielle Thierry-Mieg*, Jean Thierry-Mieg*, Ting Han*, Sebastian D. Mackowiak, Emily Mis, Charles Zegar, Michelle R. Gutwein, Vishal Khivansara, Oliver Attie, Kevin Chen, Kourosh Salehi-Ashtiani, Marc Vidal, Timothy T. Harkins, Pascal Bouffard, Yutaka Suzuki, Sumio Sugano, Yuji Kohara, Nikolaus Rajewsky, Fabio Piano, Kristin C. Gunsalus, John K. Kim (* denotes equal contribution).

microRNA target sites, and polyadenylation of most histone genes. Our data reveal a rich complexity of 3′UTRs, both genome-wide and throughout development.

4.2 Introduction

The 3'UTRs of mRNAs contain *cis*-acting sequences that interact with RNA binding proteins and/or small non-coding RNAs (*e.g.* miRNAs) to influence mRNA stability, localization, and translational efficiency (18, 172, 173). The differential processing of mRNA 3' ends has demonstrated roles in development, metabolism, and disease (174, 175). Despite these critical roles, genome-wide characterization of 3'UTRs lags far behind that of coding sequences (CDSs). Even in the well-annotated genome of *C. elegans*, nearly half (~47%) of the 20,191 genes annotated in WormBase (release WS190) (176) lack an annotated 3'UTR, and only ~1,180 (~5%) of genes are annotated with respect to alternative 3'UTR isoforms (Figure 4.5, A and B).

The sequencing of several metazoan transcriptomes and the re-annotation of expressed sequence tags (ESTs) in the mouse and human genomes have yielded key insights into 3'UTR diversity in these organisms (177-182). Complementing genome-scale analyses, directed studies in mammals have revealed remarkable 3'UTR length heterogeneity during embryogenesis (183, 184), in proliferating cells (185), and in different cellular tissues (186), suggesting that the usage of alternative 3'UTR isoforms is a highly regulated process. 3'UTR heterogeneity is due largely to the differential use of PAS motifs, resulting in

alternative 3'UTRs with different 3' end coordinates. The most conserved PAS is the 5'-AAUAAA-3' hexamer, which typically is located 19 nucleotides (nt) upstream of the polyA addition site (187). The AAUAAA hexamer is highly conserved among metazoans and binds to a family of cleavage and polyadenylation specificity factors (CPSFs) to induce the processing of the 3' ends of nascent mRNAs (188, 189).

4.3 Results

We have taken a multifaceted, empirical approach to defining the 3′UTR landscape in *C. elegans* (Figure 4.6 to Figure 4.9 and Table 4.4 to Table 4.4). We prepared developmentally staged cDNA libraries composed of mostly full-length clones spanning from 5′ capped first base to polyadenylated (polyA) tail, and we annotated 16,659 polyA addition sites in 11,180 genes by manually curating ~300,000 Sanger capillary sequence traces in National Center for Biotechnology Information (NCBI) AceView (178). We developed a method to capture the 3′ ends of polyadenylated transcripts genome-wide by deep sampling and generated a comprehensive developmental profile comprising more than 2.5 million sequence reads from Roche/454 (Figure 4.6 to Figure 4.9 and Table 4.4 to Table 4.4). We cloned 3′ rapid amplification of cDNA ends (RACE) products directly targeting 3′UTRs for 7105 CDSs (6741 genes) in both the Promoterome (190) and ORFeome (191) collections, and we recovered one or more sequenced isoforms for 85% of the targets (Figure 4.6 and Figure 4.9 and Table

4.4 to Table 4.4) (192). Finally, we remapped and annotated polyA addition sites in published RNA-seq data (182, 193).

All data sets were mapped, cross-validated, consolidated, and filtered to eliminate obvious experimental artifacts, including internal priming on A-rich stretches (Figure 2.1A). These data sets are not yet saturated: Whereas for most genes (11,516 or 73%), at least one 3′UTR isoform is supported by two or more experimental approaches, 47% of transcripts are observed by only one method (in part due to limitations specific to each protocol) (Figure 4.1 and Table 4.3 and Table 4.4). The resulting 130,090 distinct polyA sites, identified at single-nucleotide resolution and supported by more than 3 million independent polyA tags, were clustered into 26,967 representative polyA sites. Due to biological variation, 86% of tags occur within 4 nucleotides of representative sites, although individual polyA tags may spread over ~20 nucleotides (Figure 4.10).

Linking polyA sites to their parent genes proved to be a challenge, as many previous gene models were incomplete or incompatible with our new data. Using all available empirical evidence, we reannotated in AceView the *C. elegans* gene models (178). Of the 15,683 protein-coding genes with both polyA sites and cDNA support, 57% confirm the structure of WormBase WS190 gene models. The remainder encode different proteins, usually representing different cDNA-supported splice patterns: ~25% share the same stop codon, ~12% use a different stop (hundreds of those correspond to fusions or splits of earlier gene models), and ~6% are not yet annotated in WormBase.

This integrated collection, herein called the 3′UTRome (Figure 4.5), provides evidence supporting 3′UTR structures for ~74% of all *C. elegans* protein-coding genes in WormBase WS190, including previously unannotated isoforms for ~7397 genes (Figure 4.5, A to D). The length distribution of 3′UTRs parallels that in WormBase (Figure 4.5D), with a mean of 211 nucleotides (nt) (median = 140 nt). The 3′UTRome matches 61% of WormBase 3′UTRs within ±10 nt (6714 polyA ends for 6563 genes) and contains thousands of longer or shorter isoforms (Figure 4.5A). We identified 6177 polyA ends for 4466 genes with no previous 3′UTR annotation and discovered 1490 polyA ends for 1031 genes not yet represented in WormBase (Figure 4.5A).

We annotate more than one 3′UTR isoform for 43% of 3′UTRome genes (Figure 4.5 and Figure 4.11). Of these, a majority (65%) reflects alternative 3′-end formation at distinct locations in the same terminal exon for proteins using the same stop; the remainder use distinct stops in the same last exon or distinct last exons. Very rarely (79 examples), an intron within the 3′UTR is excised or retained (Figure 4.12), potentially affecting functional sequence content elements (Figure 4.12C). Indeed, putative binding sites for miRNAs (this study) or ALG-1 (15) were identified in the variable regions of some of these transcripts. About 2% of genes possess five or more 3′UTR isoforms (Figure 4.1A, Figure 4.5B, and Figure 4.11).

To identify putative cis-acting sequences that may play a role in 3′-end formation, we scanned the 50 nt upstream of the cleavage and polyA addition sites for all possible 5- to 10-mers and assigned the most likely polyadenylation

signal (PAS) motif to each 3′UTR using an iterative procedure based on enrichment and centering of the k-mers. The canonical PAS motif AAUAAA (seen in 39% of 3′ ends) and many variants differing by 1 to 2 nt are detected, with distributions all peaking 19 nt upstream of the polyA site (Figure 4.13, Figure 4.14, and Table 4.5). The canonical signal predominates in genes with unique 3′UTRs (57%). However, many high-quality 3′UTRs (3658) lack a detectable PAS motif altogether (Figure 4.1, B and C). All PAS variants are embedded within a T-rich region that spikes 5 nt downstream of the PAS motif and extends about 20 nt beyond the cleavage site (Figure 4.1D). 3′UTRs with no PAS tend to be T-rich throughout, except for a very A-rich eight-nucleotide region just after the cleavage site (Figure 4.1D). Thus, a functional PAS motif with strict sequence specificity appears dispensable for 3′-end formation in *C. elegans*.

Among genes with alternative 3′UTRs, successive polyA sites show a marked asymmetry: The longest isoform prefers a PAS, whereas shorter isoforms more often show no PAS (Figure 4.1C and Figure 4.15). The distance between alternative polyA sites peaks at ~40 nt, with resonances at ~80 and ~140 nt (Figure 4.15A). This regularity suggests that a physical constraint (possibly queuing transcription complexes) could contribute to cleavage and polyA addition at some upstream sites, which may, therefore, depend less on instructive cues from signal sequences.

Because many *C. elegans* genes undergo trans-splicing of a splice leader (SL) to the 5′ end of a nascent transcript (194), we asked whether any properties of transcript 5′ and 3′ ends correlate (Figure 4.2, A and B). About 15% of *C.*

elegans genes belong to transcriptional units called operons, each containing two to eight genes that can be cotranscribed, cleaved into separate transcripts, polyadenylated, and trans-spliced with specific leaders (Figure 4.2, A and B). The first gene in an operon is trans-spliced only to SL1; downstream genes are usually trans-spliced to 1 of 11 other SLs (SL2 to SL12), although we observed that two-thirds of these genes occasionally become trans-spliced to SL1. The processing of adjacent operon transcript ends (cleavage, polyA addition to the upstream transcript, and SL addition to the downstream transcript) is coupled mechanistically by machinery resembling the cis-splicing apparatus (195). Comparing 3′UTRs within operons, we observe that the “first” (SL1-spliced), “middle” (any gene between first and last), and “last” genes progressively decrease in average length (from 266 to 213 nt), number of 3′UTR isoforms per gene (from 2.64 to 2.51), and frequency of 3′UTRs with no PAS (from 23 to 18% in ~1400 sites) (Figure 4.2B).

However, only a small fraction (13%) of the 7026 mainly SL1-spliced genes clearly belongs to an operon, and these genes differ notably from non-operon SL1-spliced genes in their usage of the canonical AAUAAA hexamer (22% of 1409 sites versus 32% of 10,879 sites, respectively). Furthermore, we observed the canonical PAS motif much more frequently in non–trans-spliced than in SL-containing transcripts (43% of 5131 sites versus 30% of 14,873 sites) (Figure 4.2A). Whereas “standard” non–trans-spliced genes have ~30% more 3′UTR isoforms per gene than “isolated” ones having no neighbor within 2 kb (2.4 versus 1.7), these non–trans-spliced genes are more similar to each other than

to trans-spliced genes, because they have shorter and fewer 3′UTR isoforms and higher canonical PAS usage. Thus, trans-splicing within operons appears to enhance (directly or indirectly) the activity of noncanonical PAS sequences upstream, and trans-splicing at the 5′ end correlates with distinct properties at the 3′ end of the same transcript, independent of 5′-end processing downstream.

Unexpectedly, the 3′UTRome reveals polyadenylated transcripts for nearly all histone genes (Figure 4.16 and Table 4.6). The major class of replication-dependent histones (H2a, H2b, H3, and H4) is not thought to be polyadenylated in metazoans; instead, their 3′ ends form a stem-loop structure that is recognized and cleaved several nucleotides downstream by U7 small nuclear ribonucleoprotein and factors such as stem-loop binding protein (196, 197). *C. elegans* has 61 cDNA-supported histone genes (178) that all harbor conserved sequences with 3′ stem-loop potential; however, they also contain conserved PAS elements downstream of the hairpin sequence (198). Because *C. elegans* histone transcripts have also been shown to terminate in the typical stem-loop structure and to be depleted in successive rounds of polyA selection (198), we were surprised to recover polyadenylated transcripts for 57 histone genes in multiple, independent data sets (Figure 4.16 and Table 4.6). This finding suggests that, at least in *C. elegans* (and perhaps also in higher metazoans), the usual route for histone mRNA 3′-end processing may include initial cleavage and polyA addition at conserved PAS sites, followed by further processing to remove sequences downstream of the stem-loop.

We searched 3′UTRs for conserved sequence motifs and other potential functional elements. We updated our atlas of predicted conserved miRNA targets for the 3′UTRome, using the PicTar algorithm with new 3- and 5-way multispecies alignments (Figure 4.3, Figure 4.17, and Table 4.7). Roughly half of the newly predicted sites match our previous predictions (199), but many sites are gained or lost (Figure 4.17A and Table 4.7). These differences reflect improvements in both 3′UTR annotations and multispecies alignments, which increase the accuracy of conserved-seed site identification and signal-to-noise ratios. More than 3000 PAS motifs are positionally conserved among Caenorhabditis species, including within alternative 3′UTRs (Figure 4.17B). Thus, maintenance of multiple specific 3′ termini may be functionally important for some genes. Thousands of unexplained conserved sequence blocks of varying lengths within 3′UTRs (Figure 4.3B and and Table 4.7) may represent previously unrecognized functional elements that await further characterization. In vivo Argonaute (ALG-1) binding sites (200) overlap significantly with predicted miRNA target sites but not with other conserved blocks (Table 4.7), indicating that the latter are, overall, not directly related to microRNA function. For 1876 convergently transcribed neighboring genes, overlapping 3′ regions could pair as double-stranded RNA if coexpressed, potentially triggering endogenous small interfering RNA production (201) that could down-regulate cognate mRNAs (Figure 4.18).

We examined alternative 3′UTR isoforms in different developmental stages (Figure 4.4) and found a downward trend in average length and number

of 3′UTRs per gene from the embryonic through the adult stage (Figure 4.4, A and B). Among genes expressed in more than one developmental stage, embryos display the largest proportion of stage-specific 3′UTR isoforms, and these tend toward longer isoforms (Figure 4.4, B and C, Table 4.8, and Table 4.9). Some genes switch 3′UTR length coincident with developmental transitions, most notably from embryo to L1, L1 to dauer entry, dauer exit to L4, and in adult hermaphrodites versus males (Figure 4.4D and Table 4.9). Thus, 3′UTR-mediated gene regulation may be widespread in the *C. elegans* embryo, and differential expression of alternative isoforms may represent a mechanism to engage or bypass 3′UTR-mediated regulatory controls in specific developmental contexts (21, 202).

The 3′UTRome compendium evidences support for multiple mechanisms of transcript 3′-end formation in *C. elegans*, including standard PAS-directed 3′-end formation from a large collection of PAS variants, regularly spaced “shadow” polyA addition sites devoid of recognizable signals, and both operon-dependent and -independent correlations between features at the 5′ and 3′ ends of the same or of consecutive transcripts that are consistent with the possibility that trans-splicing and 3′-end processing within a gene could occur by functionally linked mechanisms. We characterize thousands of previously unknown and alternative 3′UTR isoforms throughout development, define a comprehensive catalog of PAS elements, discover a surprising number of polyadenylated transcripts with no discernable PAS, and definitively document polyadenylation of histone transcripts. We also identify conserved sequence elements in 3′UTRs that may

interact with trans-acting factors such as miRNAs and RNA-binding proteins, some of which occur within variable regions of alternative 3′UTRs. A collection of cloned 3′UTRs for several thousand *C. elegans* genes is available to the research community for high-throughput downstream analyses and in vivo studies (Table 4.10).

4.4 Supplementary Materials and Methods

4.4.1 PolyA capture

Strains. Worms were grown on NGM plates seeded with *E. coli* OP50 to adulthood. For collection of staged samples, the wild-type N2 strain was used. Embryos were isolated from gravid worms by standard alkaline/hypochloride treatment. A sample of embryos was frozen down in TriReagent (Ambion, Austin, TX), and the remainder hatched overnight in M9 buffer to yield synchronized L1 stage worms. Starved L1 larvae were plated and fed on NGM plates seeded with OP50 *E. coli* and raised at 20°C. Synchronized staged samples were collected at ~8 hr (L1), ~20 hr (L2), ~30 hr (L3), ~45 hr (L4), and ~70 hr (adult hermaphrodite). The developmental stage of each sample was verified by monitoring the seam cell lineage using Nomarski optics (Olympus, Center Valley, PA). For adult male isolation, the CB1489 *him-8(e1489)* strain was used, which increases the percentage of XO males to ~37% of the population versus ~0.2% males in the N2 wild-type strain. The *him-8(e1489)* embryos were synchronized by bleaching and incubated overnight at room temperature. Hatched L1s were aliquoted onto NGM plates seeded with *E. coli* OP50 and grown at 20°C for 4

days. Male adults were isolated by filtering through 35 μm nylon mesh, resulting in >95% males in the final sample. For dauer larvae preparation, CB1370 *daf-2 (e1370)*, CB1372 *daf-7 (e1372)*, DR47 *daf-11 (m47)*, DR2281 *daf-9 (m540)* mutants from starved plates were collected, resuspended in M9 buffer containing 1% SDS, and incubated for 20 min at room temperature. The suspension was then washed with M9 buffer and worms were placed on a fresh unseeded plate at 20°C for 12 h. Live worms that had crawled away from the dead worms were collected as dauer larvae. Worms were washed off plates with M9, washed 5 times with M9 to remove residual bacteria, and frozen in TriReagent.

RNA preparation. Total RNA was extracted using TriReagent following the vendor's protocol with the following modification: three freeze-thaw cycles (freeze in liquid nitrogen / thaw at room temperature / vortex 1 min) were included to increase worm lysis efficiency; RNA was precipitated with isopropanol at -80°C for one hour. To subtract 72 most abundant ribosome subunit genes, 25μg total RNAs were mixed with antisense DNA oligos (IDT, Coralville, IA) targeting the last DpnII site of each of these genes and digested with RNaseH (Invitrogen, Carlsbad, CA), which only cleaves RNA in RNA:DNA duplex. After subtraction, PolyA+-selected mRNAs were isolated from total RNA using oligo(dT) magnetic beads (Invitrogen, Carlsbad, CA) using the manufacturer's protocol.

cDNA synthesis. First-strand synthesis was carried out using Superscript III reverse transcription kit (Invitrogen, Carlsbad, CA) with ~20 ng of PolyA+-selected mRNA and 10 pmol of biotinylated reverse primer at 50°C for 30 min

followed by incubation at 42°C for 30 min. The following biotin-labeled primer was synthesized by Integrated DNA Technologies (Coralville, IA) and PAGE-purified: 5'Biotin-*TAATAC*- GGCGCGCCGCCTTGCCAGCCCGCTCAG-T_{20}-VN-3'. The poly(dT) and two nucleotide anchor (VN) target the proximal end of the mRNA polyA tail. The second strand was synthesized using DNA polymerase I in the presence of RNase H for 2.5 hr. The double-stranded cDNA product was extracted twice with 200 µL phenol/chloroform/ isoamyl alcohol (25:24:1), ethanol precipitated, and dissolved in 20 µL H_2O.

DpnII digestion. The resulting cDNA was digested with *DpnII* restriction enzyme (New England Biolabs, Ipswich, MA) at 37°C for 1 hr, extracted twice with 200 µL phenol/chloroform/isoamyl alcohol (25:24:1), and then ethanol precipitated and dissolved in 20 µL H_2O.

Binding biotinylated cDNA to magnetic beads. 100 µL of Streptavidin-Dynabeads M-280 (Invitrogen, Carlsbad, CA) were prepared in a 1.5 mL Eppendorf tube and then washed twice with 1 mL TE (10mM Tris-HCl, PH7.5, 1mM EDTA) and twice with 200 µL 1X B&W buffer (5mM Tris-HCl, PH7.5, 0.5mM EDTA, 1M NaCl). The beads were resuspended in 100 µL 2X B&W buffer (10mM Tris-HCl, PH7.5, 1mM EDTA, 2M NaCl). 10 µL of *DpnII*-digested cDNA fragments and 90 µL H_2O were added to the beads. The tube was rotated for 30 min at room temperature and then the beads were washed twice with 200 µL 1X B&W buffer and twice with 200 µL TE.

Ligation of barcoded linkers to the bound cDNA. Immediately after binding to Dynabeads, cDNAs were ligated to 5 µL Linker A (10 µM) using T4

DNA ligase (Invitrogen, Carlsbad, CA) (5 U/μL) for 2 hr at 16°C with intermittent gentle mixing. The beads were washed twice with 200 μL 1X B&W buffer, washed twice with 200 μL TE, and resuspended in 200 μL TE. Linker A was prepared by annealing the following two complementary oligonucleotides in TE plus 50 mM NaCl: 5'-GCCT-CCCTCGCGCCATCAG-XXXX-3' and 5'-phosphate-*GATC*-XXXX-CTGATGGCGCGAG GGAGGC-3', where *GATC* is the *DpnII* restriction sequence and XXXX represents a four-base barcode tag specific to each developmental stage: CATG (embryo), TAGT (L1), GATC (L2), CACT (L3), TACG (L4), or GAGC (adult hermaphrodite).

3' cDNA recovery. 100 μL beads were mixed with 100 μL phenol/chloroform/isoamyl alcohol (25:24:1), incubated at 65°C for 30min, vortexed at full speed for 5min, and centrifuged at 15,000 rpm for 5 min. The supernatant was collected using Phase Lock Gel (5PRIME Inc., Gaithersberg, MD). DNA was ethanol precipitated and resuspended in 20 μL H_2O.

PCR amplification. The ligation products from each developmental stage were used as template for two sequential rounds of PCR using 1 μL of DNA, the forward primer set 5'-GCCT-CCCTCGCGCCATCAG-XXXX-3', and the reverse primer set 5'- GCCTTGCCAGCCCGCTCAG-X-TTTT-X-TTTT-X-TTTT-X-TTTT-3', where the four Xs represent the four nucleotides of the stage-specific barcode tag distributed in order along a polyA tail. The periodic insertion of the X nucleotides improves reliability of Roche/454 sequencing by decreasing homopolymerization of Ts. Samples were extracted with

phenol/chloroform/isoamyl alcohol (25:24:1), ethanol precipitated, and resuspended in 50 µL H_2O. DNA concentration was measured using a Nanodrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE).

454 GS FLX Sequencing. Deep sequencing was performed on the Genome Sequencer FLX system (Roche/454 Life Sciences, Branford, CT) following the manufacturer's protocol.

4.4.2 3'RACE

3'RACE RNA extraction. Total RNA from *C. elegans* N2 mixed developmental stages was prepared using an adaptation of the RNeasy Mini kit (Qiagen, Valencia, CA). Worms were grown on NGM plates seeded with *E. coli* OP50, washed with M9 buffer, transferred to an RNase-free Eppendorf tube, and dipped into liquid nitrogen. Worms were ground using RNase-free pestles and incubated with RLT buffer (Qiagen) and beta-mercaptoethanol. The lysate was homogenized by aspiration through a 20-gauge needle fitted to a syringe and centrifuged at 13,000 rpm for 3 min. The supernatant was transferred to RNAse-free tubes and treated as per the manufacturer's recommendations.

Primer Design. Forward primers were designed to target 7,077 CDS-specific regions from WormBase WS150 for CDSs also contained in the Promoterome (190) and the ORFeome (191, 203) collections. For each CDS, in-frame sequence just upstream of and including the STOP codon (based on spliced transcript models) was selected to achieve a T_m of 60°C ± 5°C during PCR amplification. Each CDS-specific sequence was preceded by the Gateway

adaptor 5'-GGGGACAGCTTTCTTGTACAAAGTGGGA-3' to allow recombination into the pDONR P2R-P3 vector (Invitrogen, Carlsbad, CA). The primer list is available at http://www.utrome.org. A universal reverse primer was used, containing a Gateway adaptor (for recombination into pDONR P2R-P3) followed by poly(dT) and a two nucleotide anchor (VN) to target the proximal end of the mRNA polyA tail: 5'-GGGGACAAACTTTGTATAATAAAGTTG-T_{20}-VN-3'. Primers were obtained from Invitrogen.

RT-PCR. Total RNA was incubated at 55°C for one hour with Superscript III reverse transcriptase (Invitrogen, Carlsbad, CA) and the universal reverse primer according to the manufacturer's specifications. PCR amplification of 3'UTRs from the single-stranded cDNA reaction was performed in 96-well plate format, using, in each well, the universal reverse primer and a different transcript-specific forward primer as follows: denaturation at 94°C for 30 sec, annealing at 60°C for 30 sec, extension at 72°C for 3 min.

Gateway BP recombination reaction and transformation. 3'UTRs were recombined into the pDONR P2R-P3 entry vector using the BP Clonase II Enzyme Mix kit (Invitrogen, Carlsbad, CA) following the manufacturer's specifications and transformed into MultiShot Stripwell TOP 10 plates (Invitrogen, Carlsbad, CA). The transformed bacteria were grown overnight at 37°C under kanamycin selection.

Sanger Sequencing. Aliquots from overnight cultures of 3'UTR minipools were used as templates for PCR with the M13 primer set as follows: denaturation at 94°C for 30 sec, annealing at 60°C for 30 sec, extension at 72°C for 3 min.

7,077 PCR amplicons were sequenced at Agencourt Bioscience Corporation (Beckman Coulter Genomics, Danvers, MA) using the ABI 3700 automated DNA sequencers.

Preparation of deconvolved 3'UTR libraries. 6,912 minipools containing 3'UTR isoforms were manually streaked onto LB kanamycin plates. From each minipool, eight single colonies were manually isolated and propagated as individual 3'UTR clonal isoforms in 96-well plates (for a total of 55, 296 colonies). Liquid aliquots of isolated clones were re-pooled into eight different super-pools using the Aquarius automated multi-channel pipetting system (Tecan Trading AG, Switzerland), resulting in eight libraries that should each contain zero (if no insert was cloned) or one unique 3'UTR isoform per targeted CDS. These deconvolved libraries (labeled A-H) were sequenced using Solexa/Illumina and FLX Roche/454 platforms.

Sample preparation and sequencing with Illumina Genome Analyzer II. Plasmid DNA was recovered using standard alkaline lysis from overnight cultures of the eight deconvolved libraries (A-H). Inserts from each library were amplified by PCR using common Forward (5'-GTTTCTCGTTCAACTTTCTTGTACAAAGTGGGA-3')and Reverse (5'-ATAATGCCAACTTTGTATAATAAAGTTGTTTTTTTTTTT-3') primers. The eight amplicon libraries were purified using MinElute columns (Qiagen), treated to create blunt ends using T4 DNA polymerase (New England Biolabs, Ipswich, MA) and T4 polynucleotide kinase (New England Biolabs), incubated overnight with DNA ligase (New England Biolabs), and then sonicated using the Bioruptor

UCD-200 (Diagenode Inc., Sparta, NJ) for 30 min in cycles of 30 sec ON, 30 sec OFF. The resulting 8 fragmented libraries were prepared for Illumina sequencing according to manufacturer's recommendations, and six of the libraries were sequenced using the Illumina Genome Analyzer II system (Illumina, Inc., San Diego, CA) in the Sachidanandam laboratory at the Mount Sinai School of Medicine (New York, NY).

Sample preparation and sequencing with 454 GS FLX. Plasmid DNA was recovered from overnight cultures of the eight deconvolved libraries (A-H) using the Wizard Plus miniprep kit (Promega, Madison, WI) and used as template for PCR amplification with eight barcode-matched primer pairs: AdaptorA::Barcode::Forward (5'- GCCTCCCTCGCGCCATCAG-XXXX-Forward-3') and AdaptorB::Barcode::Reverse (5'- GCCTTGCCAGCCCGCTCAG-XXXX-Reverse-3'), where Forward and Reverse are the same sequences used for Illumina above and barcode tags, XXXX, for libraries A-H are A: CATG, B: TAGT, C: GATC, D: CACT, E: TACG, F: GAGC, G: CTGC, H: ATCG. Barcoded PCR amplicons from all eight libraries were combined and purified using the MinElute PCR purification kit (Qiagen). Because the FLX platform output for samples of variable length is biased toward shorter reads, the combined sample was split into two equal batches: (i) untreated, and (ii) treated with the Agencourt AMPure SPRI PCR purification kit (Beckman Coulter Genomics) to enrich for longer fragments by removing fragments shorter than 100 bp. AMPure library DNA was evaluated for quality and quantified using a BioAnalyzer DNA 1000 lab chip (Agilent, Santa Clara, CA). DNA concentration in ng/μl was converted to

molecules/µl and adjusted to $2x10_5$ molecules/µl in TE buffer. The resulting fragments were prepared for 454 sequencing according to the manufacturer's recommendations and sequenced using the Genome Sequencer FLX system.

4.4.3 cDNA libraries

Two sets of polyA+-selected cDNA libraries from the Kohara laboratory and prepared from various stages of *C. elegans* development were used (totaling 152,000 cDNA clones).

First, lambda-zap embryonic and *him-8* mixed stage libraries were prepared without any amplification or rationalization steps. These libraries are of very high quality, with ~10^{-4} mismatches per base relative to the genome (after removal of ~200 errors detected in the genome) and less than 3% structural defects or artifacts.

The second set consists of full-length L1, L2, L4 and mixed stage libraries prepared by S. Sugano Y. Suzuki and Y. Kohara using the oligo cap selection procedure (204). These libraries were designed to include the entire transcript, from 5' capped first base to polyA, and are validated by the fact that >99% of the clones with a *trans*- spliced leader in this collection contain the entire leader sequence (21 to 23 bases long). These collections allowed identification of 12 varieties of SL as well as 3,953 genes that are not *trans*-spliced.

Sequencing traces from a polyA+-selected library (n=14,811 cDNA clones), generously provided by Exelixis Inc. (San Francisco, CA), along other publicly available cDNAs and EST data obtained from the NCBI Trace and

dbEST archives (in the form of either sequences or traces), were also manually curated at NCBI as part of the experimentally supported worm transcriptome project known as AceView (178).

The combined cDNA dataset provides experimental evidence for 16,659 distinct polyA sites in 11,180 genes. These data are all publicly available from http://www.aceview.org and http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly.

4.4.4 RNA-Seq datasets

Illumina data for staged samples (L2, L3, and L4 larvae and young adults) from the modENCODE transcriptome project, described in (182), were obtained from NCBI GEO (SRX001872-SRX001875). Additional published Roche/454 datasets for the L1 stage (193) were also analyzed. Together, these data provide support for 8,332 polyA sites for 7,461 genes.

4.4.5 Sequence analysis of primary datasets

Genome version. All data were aligned to *C. elegans* genome sequence version CE6 (on which WormBase WS190 gene annotations are also anchored).

PolyA capture libraries. 454 sequence data from three independent runs were pooled. Runs A and B (Run 1) comprised sequences from combined staged samples (Run A: embryo, L1-L4, adult hermaphrodite; Run B: embryo, L1-L4, adult hermaphrodite, adult male); Run C (Run 2) contained mixed sequences from four dauer mutants: *daf-2*, *daf-7*, *daf-9*, and *daf-11* (see Table 4.2 for read counts from each run). Forward reads were identified by the pattern 5'-XXXX-

GATC-N_m-X´-AAAA-X´-AAAA-X´- AAAA-X´-AAAA-3', where *GATC* is the *DpnII* restriction site, N_m is a sequence of length *m* extending from the *DpnII* site to the end of the 3'UTR, and X´X´X´X´ is the reverse complement of the matching 3'end barcode. Reads that did not contain a decipherable barcode tag were discarded. Barcodes were used to identify the library of origin for the remaining reads, and sequences were processed to remove the 5' and 3' adaptor sequences and barcode tags. Sequences retaining length ≥15 nt were aligned to the genome using BLAT (205), with a maximum intron size of 1000, minimum window size of 5, and maximum gap of 6. Best matches were selected, and multiple alignments reported if present in more than one genomic location. Alignments in PSL format were converted to SAM format using the psl2sam.pl script provided with SAMtools (206). Alignments for sequences that did not reach the polyA were set aside; the remaining alignments were further annotated.

3'RACE. RACE clones were sequenced by three different methods. Sequences from ABI or SCF files were trimmed of vector sequence and filtered for empty vectors and putative primer-dimer products. The remaining sequences were aligned to the genome using BLAT (205) and WU-BLAST 2.0 (207). Aligned regions were scanned for the presence of detectable CDS-specific primer and terminal polyA sequences (defined as 10 or more consecutive As with zero or one intervening nucleotide).

For Illumina data, 50 million sequence reads from six independently sequenced libraries were aligned to both the genome and to AceView transcripts using the AceView aligner (http://www.ncbi.nlm.nih.gov/

IEB/Research/Acembly/Software). PolyA sites were identified by trimming reads beginning with at least 5 consecutive T's or ending in at least 5 consecutive A's, and then mapping either the full remaining tag sequence or a version lacking the last two nucleotides upstream of the polyA (since we had previously determined that the cloned RACE products contained a high proportion of T to C base changes at these positions, which pair with the two anchor nucleotides in the universal reverse primer). Overlapping mapped reads were assembled into contigs, and these were used for further annotation.

From the two 454 runs, a total of ~170,000 reads corresponding to ~85,000 unique sequences were produced. Initial processing, BLAT alignment, and conversion to SAM format were the same as described above for polyA capture data.

Alignments from all three platforms were then considered together and, where possible, alignments were assigned to the putative plate-well of origin based on the identity of the corresponding primer; for deconvoluted libraries, the combination of primer and barcode, if detectable, was used to assign a putative location in the isolated clone library plates.

cDNAs and ESTs. cDNA clones from the yk collection were sequenced using the Sanger method. All cDNA and EST data from this collection and from other sources (as described above) were aligned to the genome and annotated using AceView tools; these were further hand-curated by visual inspection of multiply aligned ABI sequence traces, where available.

RNA-seq. Illumina and Roche/454 datasets (described above) were aligned to both the genome and AceView transcripts using the AceView software tools (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/Software). PolyA sites were identified trimming reads beginning with at least 5 consecutive T's or ending in at least 5 consecutive A's, and mapping the remaining sequence tag as above. Overlapping mapped reads were assembled into contigs, and these were used for further annotation.

4.4.6 cDNA and transcriptome annotation

Annotation of independent datasets. Sequences from RACE and polyA capture with best-hit alignments or assembled contigs near the last exon of a (targeted, for RACE clones) CDS were defined as candidate 3'USTs (UTR Sequence Tags). USTs were initially assigned to the overlapping or immediately adjacent upstream CDSs from WormBase WS190 gene models (http://www.wormbase.org); these assignments were later revised using AceView genes (http://www.aceview.org), which in some cases revealed that the combined data were incompatible with existing WS190 (or WS150) CDS models. In such cases, USTs from RACE experiments were retained as evidence of transcriptional activity but were removed from the final list of cloned 3'UTRs. USTs with a contiguous BLAT alignment extending through the STOP codon of a valid AceView CDS model and containing polyA sequences were considered to be bona fide complete 3'UTR isoforms with full-length coverage. Those with incomplete 3'UTR coverage and/or no detectable polyA sequence were

annotated as partial 3'USTs and used to refine 3'UTR boundaries. Mapped tags from short read data were assembled into contigs and used together with cDNA, EST, UST data to define transcribed regions. The combined data were used to refine and extend existing AceView genes. Data mapping downstream of (but not overlapping) an existing gene were extended *in silico*, where possible, and assigned to the nearest gene upstream or else used to define new transcriptional units. All annotated 3'USTs and 3'UTRs were used for subsequent analyses.

Definition of representative polyA sites and 3'UTR isoforms. To define 3'UTR isoforms and assign a single representative polyA site per isoform, we combined evidence for polyA addition sites from all four independent data sources in the 3'UTR compendium into a single large dataset and performed an iterative local clustering procedure using their chromosomal coordinates. The clustering software is included in the AceView software, available from http://www.ncbi.nlm.nih.gov/IEB/Research/ Acembly/Software. When evidence sources were attached to a known gene model, clustering was performed on a per-gene basis. The local maximum for each cluster was computed and used as the position of the reported ("representative") polyA addition site for each 3'UTR isoform. The spread of the clusters extends from one up to around 20 nucleotides, with 86% of all individual data points falling within 4 nt of the representative polyA site (Figure 4.10).

Using this clustering procedure, each 3'UTR isoform was then defined as a unique sequence span that extends from a specific CDS end and terminates downstream at a distinct "canonical" polyA addition site: 3'UTR sequences that

share the same CDS end and terminate within the same polyA cluster were defined as examples of the same isoform, whereas 3'UTR sequences that terminate within different polyA clusters (even if linked to the same CDS) were defined as distinct isoforms. Isoforms of a gene that were represented by less than 5% of the total polyA counts for that gene, isoforms that were not supported by two or more independent pieces of evidence, and those that were shorter than 20 nt (which mostly contained dubious cloning artifacts) were removed from the final dataset. For reporting purposes and all downstream analyses involving isoforms, we considered only the "representative" polyA coordinate for each reported 3'UTR isoform.

Identification of PAS sites. The 50 nt regions immediately upstream of all polyA sites were scanned in an unbiased way for all possible 5 to 10-mer sequences to identify any statistically over-represented motifs. The only motifs returned from this exercise were the canonical PAS sequence (AAUAAA) and several closely related sequences. The distribution of all over-represented hexamers peaked at a start position of -19 nt from the polyA site, which was taken as the most likely position of the PAS site. All of the 3'UTR isoforms in the compendium were then scanned for the canonical PAS sequence and any hexamer with an edit distance of 1 or 2 nt. Because it is not possible to definitively identify the "real" PAS site, we scanned for hexamers in a preferred order based on their observed frequency of occurrence in 3'UTRs between 10 and 30 nt upstream of the polyA site, and considered those occurring at a frequency of ≥1% as putative PAS motifs. We used the first occurrence of a

putative motif in the ordered list as the most likely functional PAS sequence. UTRs that did not contain one of the resulting 26 putative PAS motifs within this interval were termed “no PAS”.

Analysis of genomic nucleotide frequencies in the 120 nt region spanning ±60nt of polyA sites showed that strongly supported PAS sites, which we consider the best candidates for recognition by CPSFs for 3’end-processing (189), also show an enrichment of T’s that peaks at +5 nt downstream of the putative PAS site (Figure 4.1D). These include nine principal motifs are: AATAAA (the canonical PAS hexamer), AATgAA, tATAAA, cATAAA, gATAAA, AtTAAA, tATgAA, AgTAAA and cATgAA (where upper-case letters are identical with the canonical hexamer, and lower-case letters indicate substitutions).

Comparison of 3’UTRome and WormBase annotations. Operon, Gene, CDS, and 3’UTR annotations for WS190 were obtained from WormBase. For comparative purposes, any 3’UTR in our compendium whose 5’end matched a WS190 CDS and whose 3’end was within 10 nt of an annotated WS190 3’UTR was considered identical; all others were labeled as “longer” or “shorter” than the WS190 3’UTR, as appropriate. 3’UTRs in our dataset that matched a WS190 CDS end but had no corresponding WS190 3’UTR were annotated as “new 3’UTRs”. 3’UTRs that did not match a WS190 gene model, but matched an alternate transcript model that could be generated from experimental data, were annotated as 3’UTRs of “new AceView genes”. These data are summarized in Figure 4.5.

Intron analysis. Gapped sequence alignments were examined for the presence of putative splice signal consensus sequences, and introns were annotated as appropriate. Numerous gapped alignments of polyA capture data spanned bona fide splice junctions but were on the opposite strand and thus contained the reverse complement of known splice consensus signals. Such alignments were observed to occur most frequently within coding regions; these were determined most likely to represent mis-priming in A-rich regions and were discarded. A subset of gapped alignments for these data contained terminal segments <10 nt; these appeared to be alignment artifacts of degraded sequence data and were also discarded. A total of 363 3'UTRs for 192 genes were determined to contain bona fide introns, based on the presence of a strongly supported CDS upstream with no evidence for another CDS that could extend into the putative 3'UTR. The 3'UTRs with an intron that could also occur internally within the CDS of an alternative isoform were not counted in this set.

Operon and SL analysis. To compare the six categories of genes analyzed in Figure 4.2, we selected a subset of trans-spliced and non-trans-spliced genes for which assignment to a unique category could be unambiguously determined. Among the SL1 trans-spliced genes, we identified 574 SL1 genes occupying the first position of an operon (genes fully supported from SL1 to polyA and separated by at most 300 bases from the next gene in cis, which is itself trans-spliced mostly to SL2) and 3,530 SL1-genes undoubtedly not in an operon (selected as followed either by another SL1-gene (n=1,749) or by a

confirmed non-trans-spliced gene (n=1781)); these two subsets were found to be indistinguishable and were merged in Figure 4.2.

Directed RT-PCR assay for retained 3'UTR introns. Total RNA was extracted from mixed-stage worms and RT-PCR was performed essentially as described above. 1 µg of total RNA from mixed-stage worms was used as template for a first strand reaction using the universal anchored poly(dT) reverse primer. PCR was performed using internal primer pairs flanking putative retained introns in the 3'UTRs of two genes: *par-5* (Forward: 5'-GAG GGA AAC CAG GAA GCT GGA AAC TAA-3'; Reverse: 5'-GAT GCT ATT GCG CAG TGT TGT ATG GAG TAT TGG) and *sams-1* (Forward: 5'-GCC ACA TCT GCT ATC GCT CAC TAA-3'; Reverse: 5'-CAA GAC AGC TCA GCG GGT AGC GGA AAC CG-3'). Products were separated on a 2% agarose gel and visualized with ethidium bromide.

Developmental stage analysis. The staged polyA capture dataset was used for this analysis, since this dataset can provide specific information on the abundance of alternative 3'ends expressed in different stages. Since the total polyA tag count differed between libraries, the total number of read counts from each stage was normalized to match the total counts in embryo, and counts for individual isoforms scaled proportionally to reflect the relative expression level in different lifestages. The number of isoforms detected per gene was evaluated for each developmental stage and across all stages. To study the expression of long vs. short isoforms we identified genes showing exactly two distinct 3'UTR isoforms (2,295 in total) and restricted our analysis to a stringent subset of 1,960

genes showing at least 5 read counts for the most abundant isoform (Supplementary Dataset S5). To identify genes showing preferential isoform usage, we further selected a subset of genes that showed, in the cumulative dataset, at least twice as many total counts for one isoform as the other (915 genes for long>short; 615 genes for short>long). The per-stage relative expression of a particular isoform of a gene was calculated by dividing the counts for that isoform by the total counts for both isoforms expressed during that stage. The relative expression of an isoform across all stages was calculated as the ratio of the normalized counts of the isoform in a single stage to the total normalized counts of both isoforms of the gene across all developmental stages.

To identify genes that exhibit a differential preference for 3'UTR isoforms during development (i.e. 3'UTR isoform "switching"), we filtered the 1,960 genes described above using the following criteria: 1) isoform *'a'* was more abundant than the isoform '*b*' in one developmental stage, and isoform '*b*' was more abundant than isoform '*a*' in any other developmental stage; 2) the total abundance of all isoforms for the same was ≥ 20 counts (abundance was based on normalized polyA capture counts). We identified 612 genes exhibiting such 3'UTR isoform switching. To obtain a "high-confidence" subset of these genes, we imposed two additional criteria: 1) the ratio of counts for isoform '*a*' to counts for isoform '*b*' (a/b) was ≥2 fold in one stage, and the ratio of isoform '*b*' to isoform '*a*' counts (b/a) was ≥2 fold in another stage; 2) the difference in support between isoform '*a*' and '*b*' was ≥5 counts within each developmental stage in which switching occurred. Of the 612 genes, 263 genes passed these filters.

miRNA target prediction and 3'UTR conservation analysis 3'UTR alignments. We used the Galaxy server processing pipeline (208) and the UCSC Table Browser (209) to prepare a multiple alignment file (MAF) for *C. elegans* (WS190/CE6), *C. remanei, C. briggsae, C. brenneri*, and *C. japonica*. The MAF file did not contain overlapping blocks or gaps in the *C. elegans* sequence. We then extracted a MAF file for each of the initial 33,909 3'UTRs from the 3'UTRome. Overlapping 3'UTRs were fused to yield 15,685 unique 3'UTR regions that were used for subsequent analyses.

miRNA sequences. We used for our analyses 174 *C. elegans* mature miRNA sequences downloaded from miRBase version 14 (210) and 9 novel miRNAs determined by miRDeep2 (211). These miRNAs were grouped into 124 miRNA families sharing the same seed sequence at nucleotides 2-7 in each miRNA.

Identification of miRNA seeds in 3'UTRs. The PicTar algorithm (199, 212) was used to identify non-conserved and conserved miRNA seeds in mRNA sequences, which were defined as regions in mRNA sequences with perfect base complementarity to miRNA 6-mer seeds (nucleotides 1-6 or 2-7 at the miRNA 5' end). Seeds conserved in 3 species (*C. elegans, C. remanei*, *C. briggsae*) and those conserved in 5 species (*C. elegans, C. remanei, C. briggsae, C. brenneri*, *C. japonica*) were identified. PicTar was further used to predict and assign scores for full miRNA binding sites, as described (199). The probability of a conserved predicted miRNA target seed site being functional in 3-way or 5-way species comparisons is 2.7 and 3.1, respectively. The

comprehensive list of PicTar predictions is available from the UTRome (http://www.utrome.org) and modENCODE (http:// www.modencode.org) websites.

Comparison with Lall *et al*., 2006. We compared our updated miRNA target predictions within our previous predictions for *C. elegans* (199). For this comparison, we considered only those miRNAs that were analyzed in Lall et al. and the set of unique (non-overlapping) 3'UTRs contained in the UTRome to which the Lall et al. target site predictions map; thus, any predicted sites from either study that were not contained in 3'UTRs considered in the other study were not included in this comparison. In addition, we excluded from the comparison the two miRNAs cel-miR-68 and cel-miR-69 used in the Lall et al. analysis (because they are currently annotated as siRNAs in WormBase), and the seven miRNAs cel-miR-42, cel-miR-239b, cel-miR-248, cel-miR-250, cel-miR- 252, cel-miR-253 and cel-miR-358 (because the reported sequences of their seed regions, i.e. positions 1-7 or 2-8 in the mature miRNA, were different according to Rfam version 6 and miRBase version 14).

We then compared the number of predicted sites from this study with the previous set of predictions within the sequence space analyzed in both studies (summarized in Table 4.7). From our new prediction set, 5,943 predicted miRNA target sites fall in this intersecting sequence space, of which 580 sites (9.8%) were not identified in the Lall et al. study. We attribute the identification of these new sites to improved multi-species alignments and the inclusion of newly sequenced species in the alignments. Of the 11,131 miRNA target sites

predicted in the Lall et al. study, 6,474 sites were located in the intersecting sequence space. In the current study, we recovered 5,363 of those sites, or 82.8%; the remaining 1,111 sites from Lall et al. (17.2%) could not be recovered. The loss of these sites is explained by the fact that the Lall et al. study used some sequence regions outside the 3'UTRome for the initial predictions; if conserved sites were identified in these regions, then non-conserved sites falling within shorter 3'UTRs would also be designated as candidate target sites due to the presence of the initial conserved site. However, if this sequence region is not used for the initial identification, and no other conserved sites are identified within the sequence space analyzed, then non-conserved sites will not be considered by the algorithm as potential target sites, and previously predicted sites would then be lost.

We note that many previously predicted target sites from Lall et al. that fall outside the spans of our 3'UTR annotations (either because they targeted genes for which we have no 3'UTR annotation, or because we previously used up to 500nt spans downstream of any CDS if no 3'UTR was available) are not currently supported by empirically defined 3'UTR regions.

Conserved blocks not explained by miRNA seeds. To identify conserved sequence blocks that do not correspond to conserved miRNA seed sequences, all (reverse complemented) miRNA seeds were masked with Ns in the 3'UTR multiple alignment files (MAFs), and all remaining k-mers ($k \geq 6$) conserved in 3 species (*C. elegans, C. remanei, C. briggsae*) or in 5 species (*C. elegans, C. remanei, C. briggsae, C. brenneri, C. japonica*) were identified. The

alignment of any conserved 6-mer was extended as far as possible in both directions.

Distribution of conserved PAS motifs and sequence blocks. We excluded from this analysis all 3'UTRs shorter than 10 nt and those contained within coding sequences of alternative CDSs, resulting in a final set of 24,858 3'UTRs, of which 8,319 genes have a single isoform, 3,320 genes have exactly two isoforms, and 2,616 have more than two isoforms. All conserved miRNA seeds in 3'UTRs, all 29 putative PAS motifs, and all conserved sequence blocks as defined above were investigated with respect to their positions relative to UTR ends. A PAS site was considered as "conserved" in this analysis if it was found in *C. elegans* and the same or another PAS motif was found within a window of ±5 nucleotides in aligned *C. briggsae* and *C. remanei* sequences. Only PAS sites in genes with one isoform or exactly two isoforms, where the longest isoform was at least 100 nt, were considered. The set of genes with 2 isoforms was further filtered to require a length difference of at least 50 nt between the short and long isoform; if this requirement was not met, the short isoform was discarded and the gene was treated as having a single long isoform for this analysis.

Analysis of overlaps between experimentally determined ALG-1 binding sites and conserved sequence motifs. We compared recently published *in vivo* Argonaute (ALG-1) binding sites (200) with our conserved sequence motifs (predicted miRNA target sites and conserved sequence blocks). For this analysis we considered only those 3'UTRs containing or overlapping at least one ALG-1 binding site. The probability of predicted miRNA target seed

sites from 3-way species alignments (*C. elegans, C. briggsae, C. remanei*) occurring within an ALG-1 binding site was 0.75. As a control, we calculated the overlap between ALG-1 sites and 6-mers (the length of predicted miRNA seed sites) placed at random positions along the length of annotated 3'UTRs (p=0.43), which represents a lower bound to the resolution at which we could discern meaningful correlations with ALG-1 sites. The overlap was not significant for the thousands of other conserved blocks that are not explained by predicted miRNA target sites or by conserved PAS sites (0.54 vs. 0.48 for random controls). These results indicate that the overlap between ALG-1 sites and predicted miRNA target sites is highly significant, and that while other conserved sequence blocks are likely functional, they are not, overall, directly related to microRNA function.

4.5 Acknowledgements

This work was supported by NIH grant U01-HG004276 to F.P., K.C.G., J.K.K., and N.R. and grants from the Muscular Dystrophy Association, NIH R01GM088565-01, and the Pew Charitable Trusts to J.K.K. The authors thank Tom Blumenthal, John V. Moran, Allison Billi, Desirea Mecenas, and Bastiaan Bergman for helpful comments and discussions on the manuscript; Heesun Shin, Shu Yi Chua, and David Baillie for providing *C. elegans* cDNA sequences; Tal Nawy for the statistical analyses; *Ravi Sachidanandam* for generous Illumina sequencing of the 3'RACE clones; Robert Lyons and Suzanne Genik for Roche/454 sequencing of polyA capture libraries during the pilot phase of the study; Philip McMenamin and Dan Schaub for help on the development of the

3’UTRome database; Mitzi Morris for help with data submission; and Lucy Huang for help on the stage analysis; and Marc Vidal, Michael Hengartner, Scott Tenenbaum, and Marv Wickens for contributions to our original modENCODE proposal.

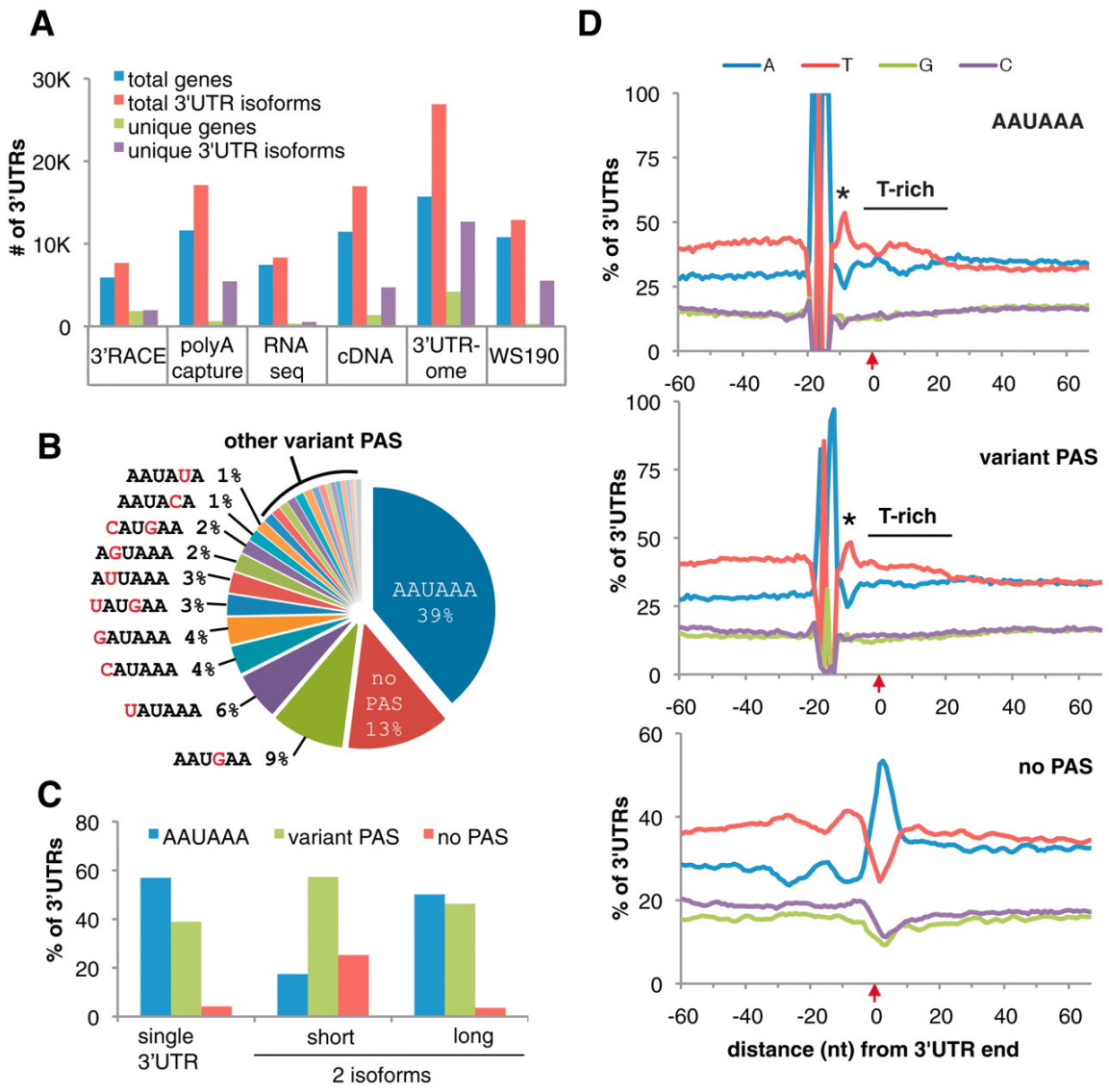


Figure 4.1 The 3'UTRome and 3'UTR PAS.
(**A**) The number of genes and isoforms detected in, or specific to, each data set and cumulative totals in WS190 and 3′UTRome annotations. (**B**) PAS motif frequencies: AAUAAA (39%), variant PAS (1 to 9%), and no PAS (13%). (**C**) PAS usage in genes with one or two (short and long) 3′UTR isoforms. (**D**) Nucleotide distribution spanning ±60 nt around the polyA addition site, in 3′UTRs with: AAUAAA (top), 10 most common variant PAS (middle), and no PAS (bottom). Alignments, centered at –19 nt, show a T-spike at 5 nt downstream of PAS (asterisk), polyA addition site (red arrow), and T-rich region downstream of cleavage site. The A-rich peak downstream of "no PAS" is not enriched for AAAAAA, suggesting an A-rich motif at that location rather than artifactual A-rich ends.

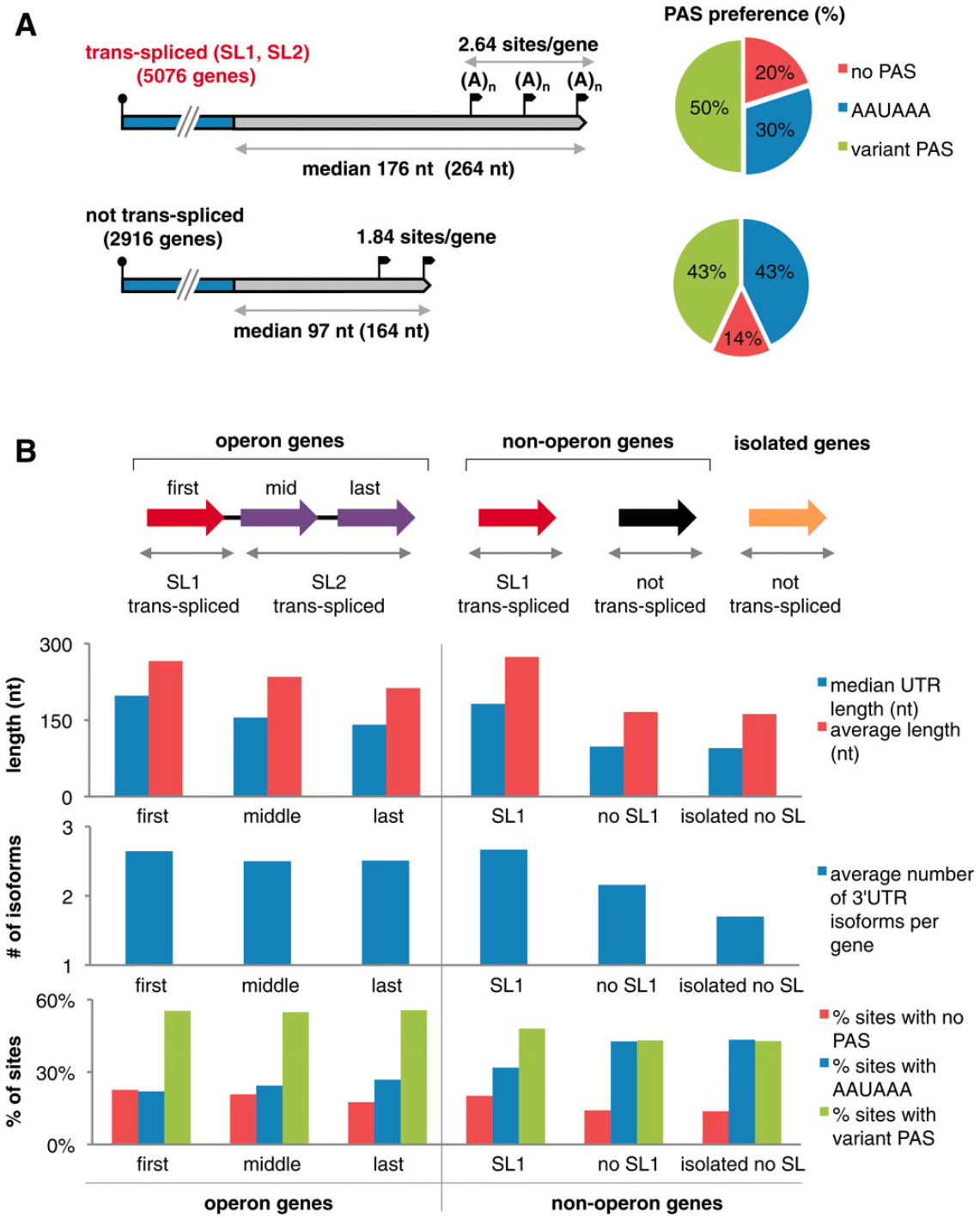


Figure 4.2 3'UTRs in operons and trans-spliced versus non-trans-spliced mRNAs.

(**A**) Trans-spliced (top) and non–trans-spliced (bottom) mRNAs: 3′UTR median (and average) lengths, number of 3′UTR isoforms per gene (polyA sites, black flags), and PAS preference (pie charts: % 3′UTRs with AAUAAA, variant PAS, and no PAS). (**B**) (Top) Schematic of operon (left, n = 574 operons), non-operon (center, n = 4348 genes), and isolated (right, n = 2098) genes. Initial operon genes (red) are SL1–trans-spliced; downstream genes (purple) usually acquire

one of the other SLs (SL2 to SL12). Non-operon genes are either SL1–trans-spliced (red, n = 3530) or not trans-spliced (black, n = 818). Isolated genes (having no neighbors within 2 kb) are not trans-spliced (orange, n = 2098). (Bottom) 3′UTR lengths, number of isoforms, and PAS sites for operon and non-operon genes.

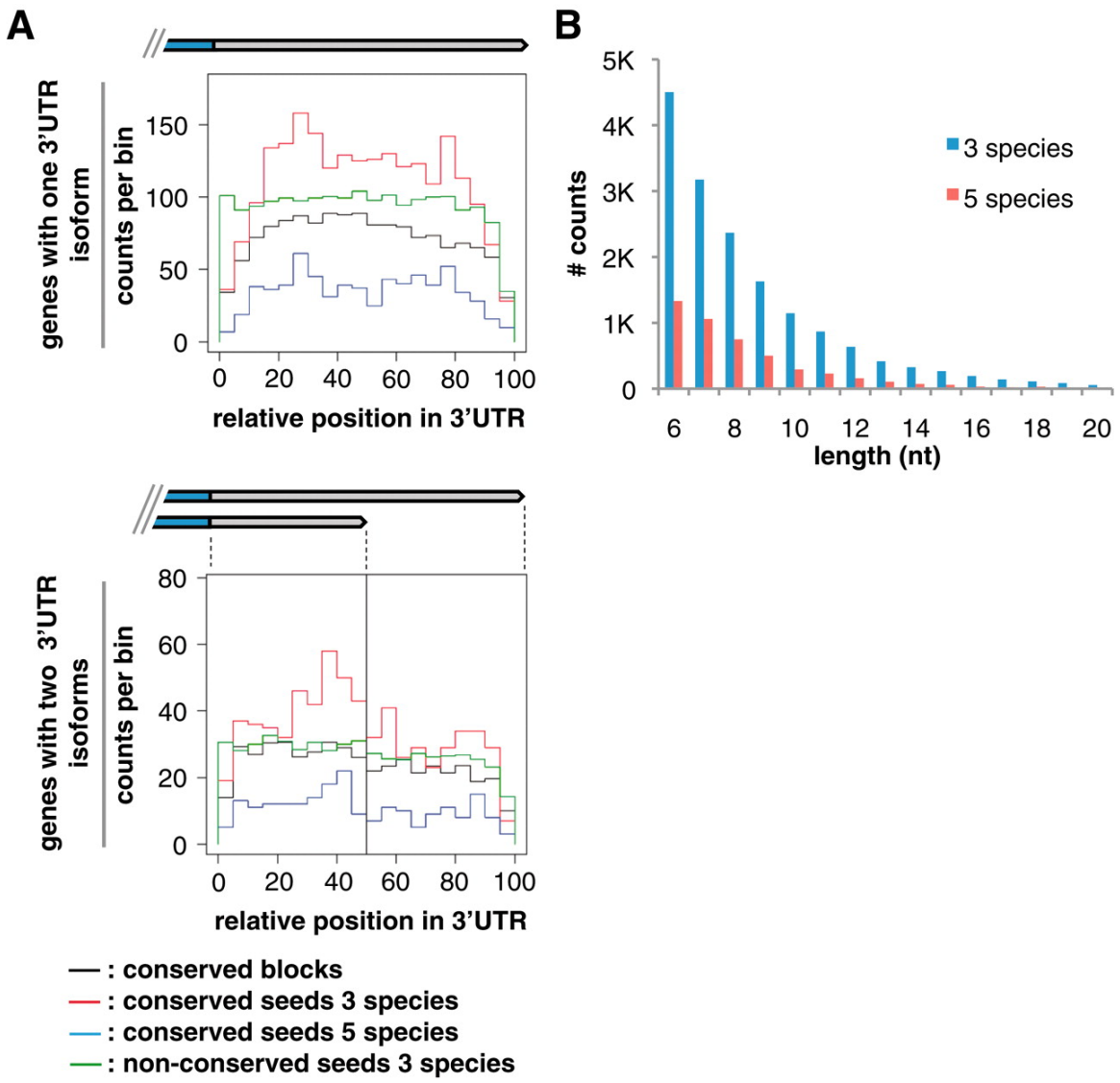


Figure 4.3 Conserved sequence elements in 3'UTRs.
(A) Histogram distributions of conserved sequence blocks (black, counts shown at 1/5th scale), conserved miRNA seeds in three (red; *C. elegans*, *C. remanei*, *C. briggsae*) and five (blue; *C. elegans*, *C. remanei*, *C. briggsae*, *C. brenneri*, *C. japonica*) species, and nonconserved miRNA seeds (green, 1/25th scale) along the normalized length of 3′UTRs, in genes with one isoform (top) or exactly two isoforms (bottom). For genes with one isoform, the length scale is 100%; for two isoforms, 0 to 50% represents the short-isoform span, and 51 to 100% indicates the span exclusive to the long isoform. Counts were binned by fraction of total length and, thus, varied in absolute length. (B) Length distribution (up to 20 nt) of conserved sequence blocks in 3′UTRs (excluding miRNA target and PAS sites), in three (blue; n = 16,204 conserved blocks) and five (red; n = 4758) species. See also Table 4.7.

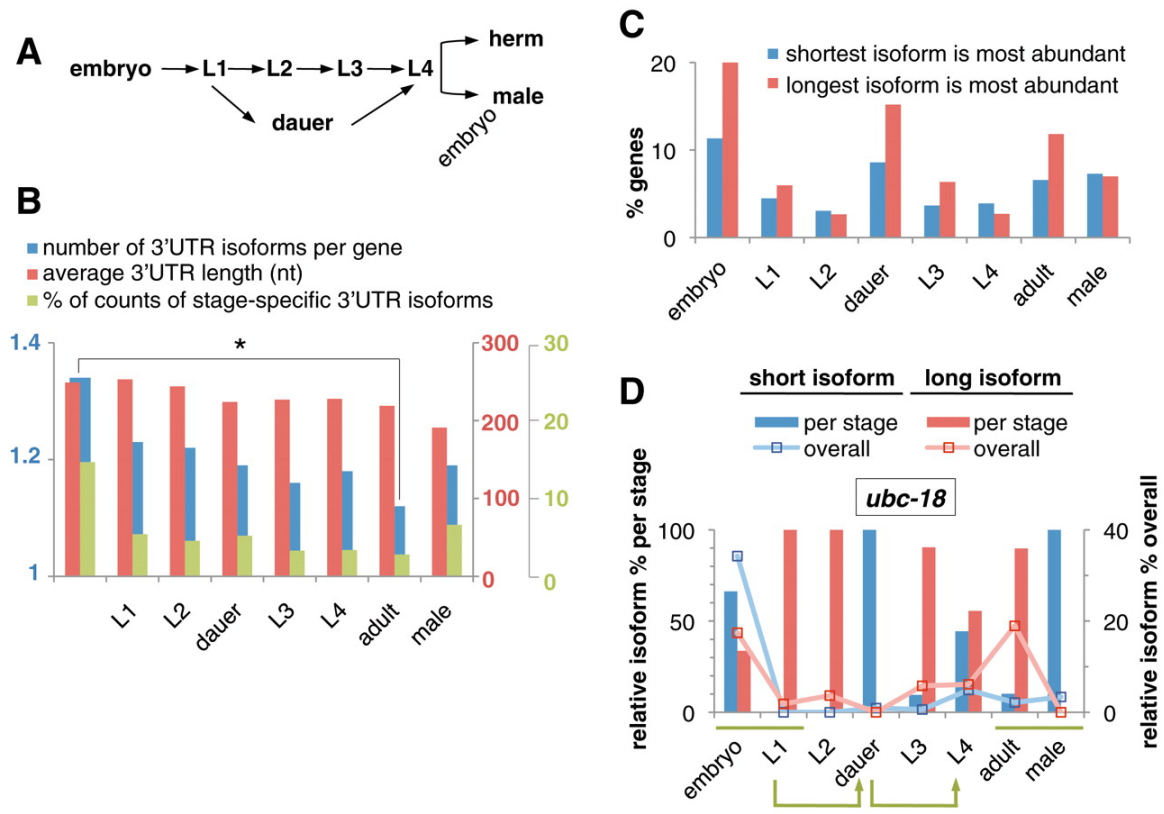


Figure 4.4 3'UTRs during development.

(A) *C. elegans* developmental transitions: embryogenesis, four larval stages, and adults. In unfavorable environments, L1 larvae arrest in dauer stage and can re-enter the life cycle as L4 larvae. herm, hermaphrodites. (B) The number of 3′UTR isoforms per gene decreases significantly during development (blue) (*p ~ 0.004, permutation test). The average length of 3′UTRs decreases during development (red). Adult males have shorter average 3′UTRs than hermaphrodites. Embryos show more stage-specific 3′UTR isoforms for genes expressed during multiple developmental stages (green) (see Table 4.8). (C) Proportion of genes showing stage-specific expression of alternative 3′UTR isoforms (see Table 4.9). Embryos and dauers favor longer 3′UTR isoforms. (D) Differential 3′UTR-isoform expression during development (*ubc-18* shown). The bar chart illustrates the relative abundance of short versus long 3′UTR isoforms for *ubc-18* in each stage (sum per stage = 100%, left y axis). The line graph shows the relative abundance across all stages (sum per gene across all stages = 100%, right y axis). Green bars highlight differences in 3′UTR isoform usage in the embryo-to-L1 transition and between adult hermaphrodite and male stages. Green arrows indicate dauer entry and exit transitions.

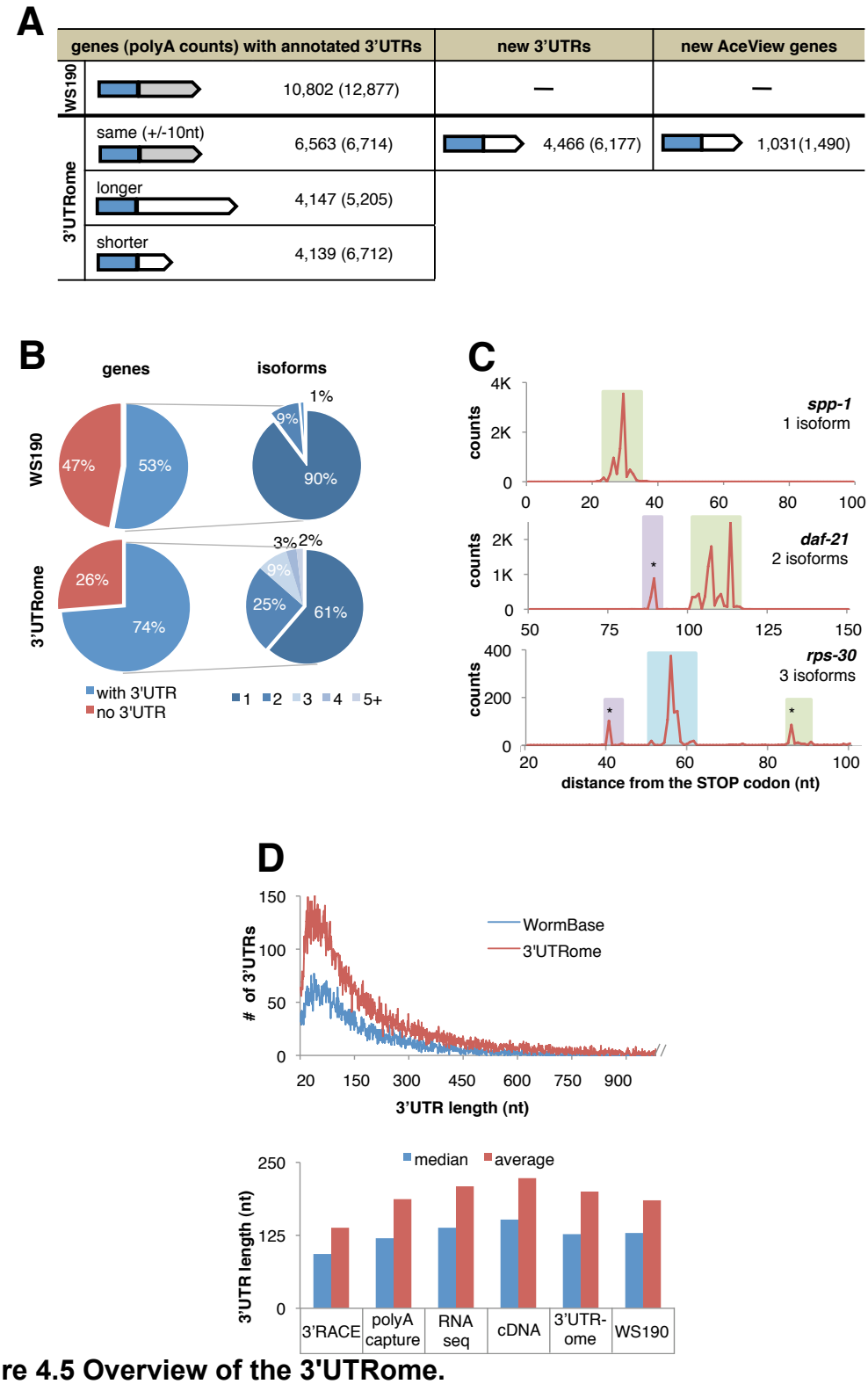


Figure 4.5 Overview of the 3'UTRome.

A,B,D) Comparison with WormBase (WS190) gene models. A) The 3'UTRome contains 3'UTRs of similar, longer, and shorter length for WS190 genes with annotated 3'UTRs (left column); 3'UTRs for WS190 genes with no annotated 3'UTRs (middle column); and 3'UTRs for transcriptional units not annotated in

WS190 (AceView genes) (right column). B) WormBase WS190 contains 3'UTR annotations for 10,802 protein coding genes (53% of total); of these, only 10% are annotated with two or more 3'UTR isoforms. Our 3'UTRome covers 14,918 WS190 coding genes (74%), 39% of which possess two or more isoforms. C) Observed counts of polyA sites from independent sequence reads cluster together, defining one or more 3'UTR isoforms. Variability within polyA clusters (colored boxes) spans up to ~20 nt. Asterisks denote newly identified 3'UTR isoforms. D) Top panel: The length distributions of 3'UTRs in WS190 and 3'UTRome datasets are homothetic. Bottom panel: median (blue bar) and average (red bar) length of 3'UTRs detected in each dataset.

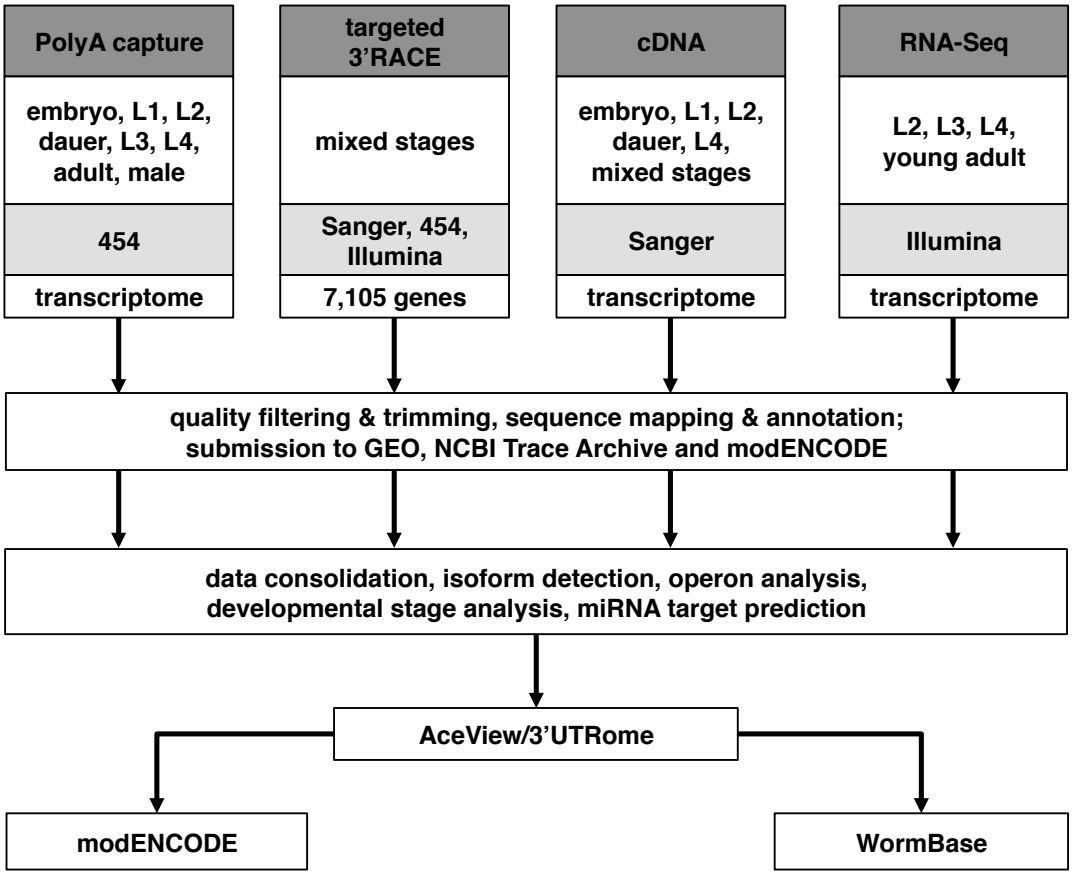


Figure 4.6 Overview of 3'UTRome pipeline.
The 3'UTRome project is composed of four datasets. PolyA capture and targeted 3'RACE were generated in this study, while publicly available cDNA and RNA-Seq data were reanalyzed and curated as part of this effort. Barcoded polyA capture tags contain the 3' end portions of 3'UTRs from staged samples; 3'RACE products directed at 7,105 coding genes were cloned from mixed stage samples. The cDNA dataset represents AceView-curated cDNA and EST sequences using, where possible, the original traces from cDNA libraries produced by the Kohara laboratory, Exelixis, and others obtained from the NCBI trace repository, as well as cDNA sequences from NCBI sequence repositories (GenBank, dbEST). The RNA-Seq dataset consists of published data for staged mRNA samples from the modENCODE *C. elegans* transcriptome project (182) and previously reported L1-stage data (193). Datasets were sequenced as indicated (gray shading). Sequences were processed (to remove vector, linker, barcode, and polyA sequences), filtered for read quality, and aligned to the *C. elegans* WS190/CE6 genome. The consolidated datasets were used to define a compendium of 3'UTR isoforms, which was used for downstream analyses of 3'UTR structure and function. Raw data and annotations for the compendium are available in public repositories, including NCBI GEO and Trace Archive, the 3'UTR-centric 3'UTRome database (http://www.utrome.org), AceView

(http://www.aceview.org), modENCODE (http://www.modencode.org), and WormBase (http://ww.wormbase.org). Supplementary Materials and Methods provide additional details on data production and analysis.

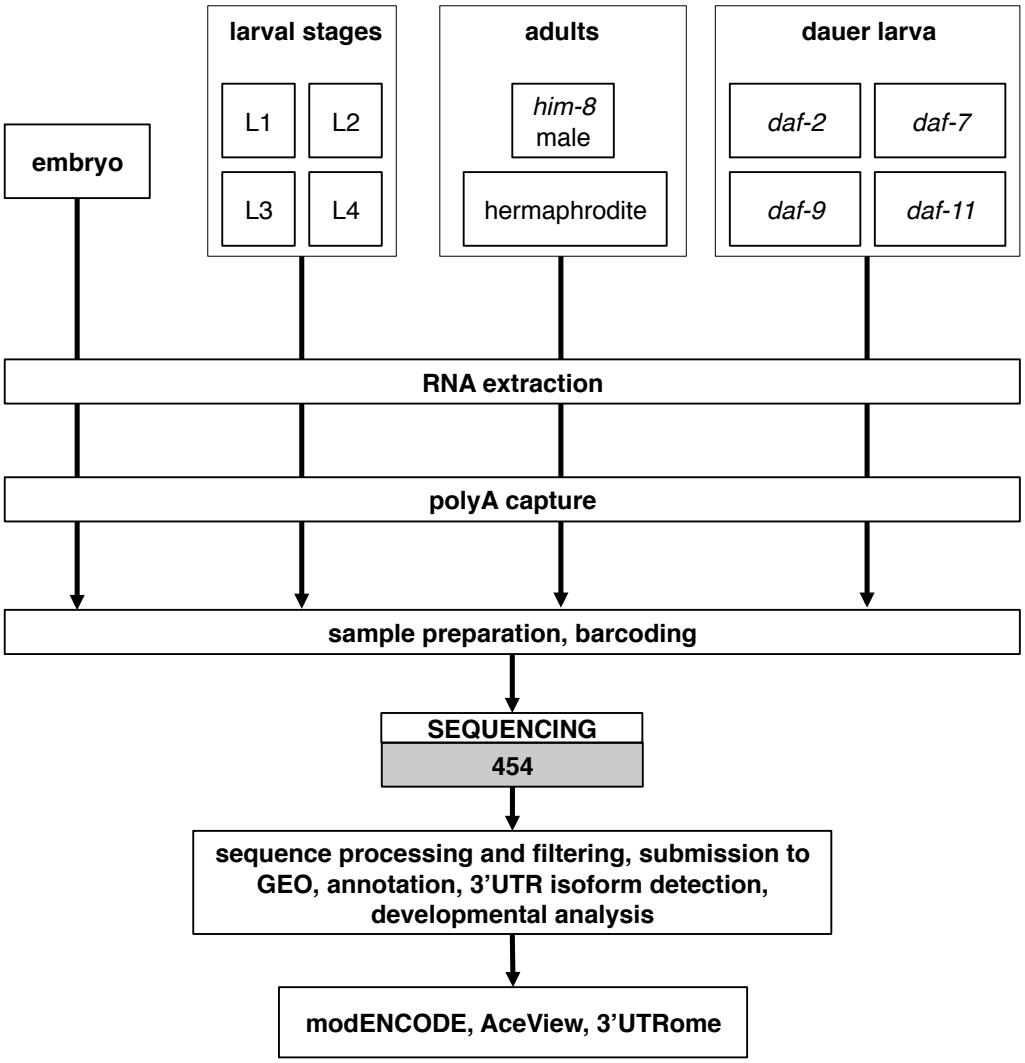


Figure 4.7 Workflow for polyA capture assay.
Barcoded polyA capture libraries were prepared using total RNA from staged animals and sequenced by Roche/454. Reads were filtered for quality, processed to remove adaptor and barcode sequences, and aligned to the WS190/CE6 genome build. Raw and processed sequence files were submitted to GEO. Alignments were consolidated with the other 3'UTR datasets and annotated with respect to WS190 and AceView gene models. Data and annotations are available in AceView, 3'UTRome, and modENCODE databases (see Supplementary Materials and Methods for details).

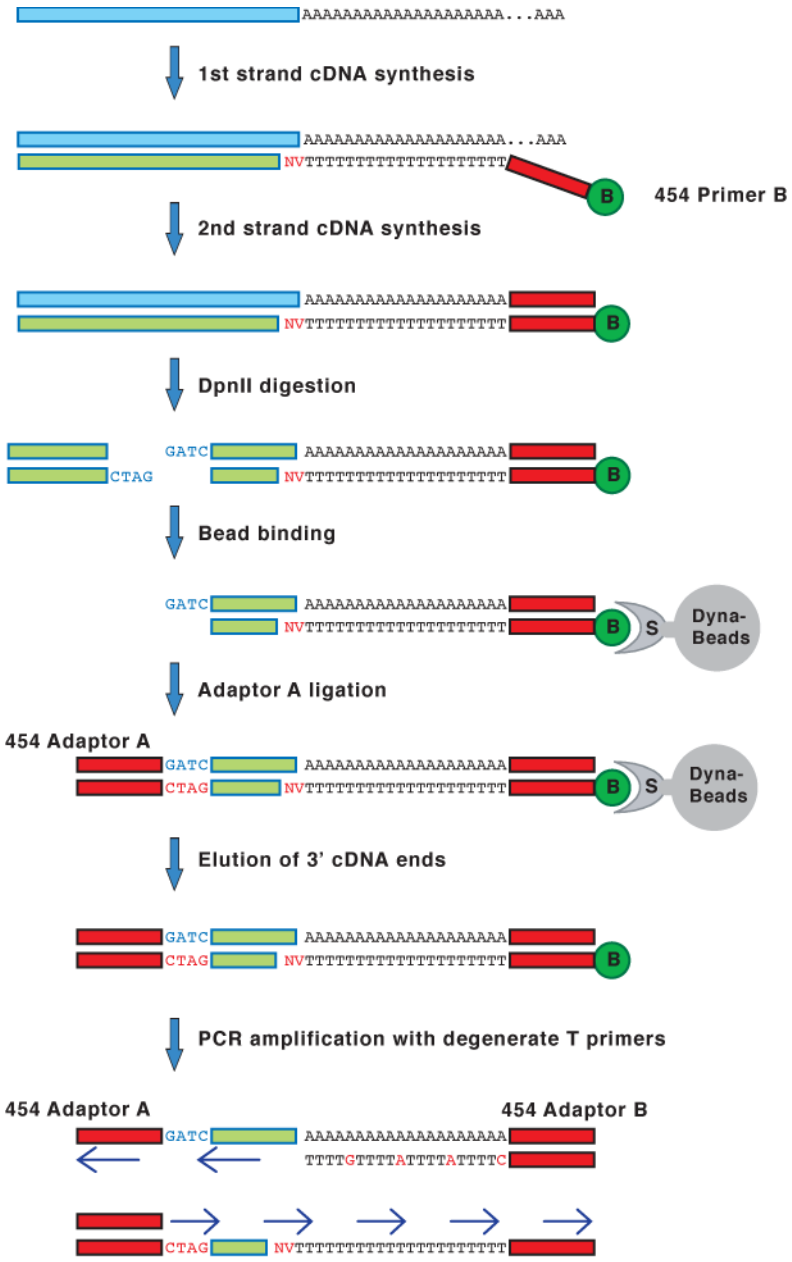


Figure 4.8 PolyA capture protocol.
Total RNA from staged samples (Figure 4.7) served as template for a first-strand reverse transcriptase (RT) reaction with an anchored, biotinylated poly-dT primer. Second-strand synthesis with T4 DNA polymerase produced dsDNA products that were digested with *DpnII*. Three-prime terminal fragments were recovered using streptavidin beads, ligated with barcoded 454 sequencing primers, PCR amplified, and subjected to pyrosequencing (see Supplementary Materials and Methods for details).

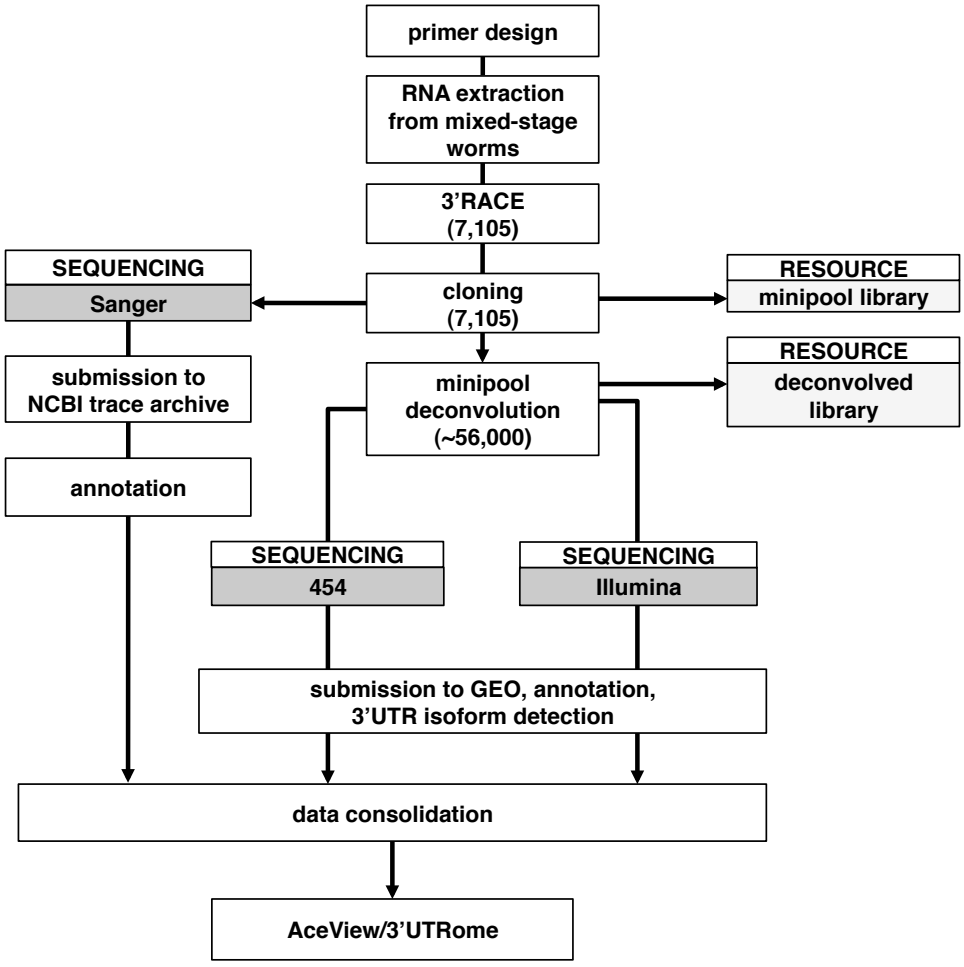


Figure 4.9 Flowwork for 3'RACE.
A 3'RACE cloning pipeline was designed to target 3'UTRs of 7,105 CDSs for 6,741 genes previously included in the Promoterome (190) and ORFeome (191, 203) collections. 3'RACE products were generated from total RNA isolated from mixed developmental stages, cloned into Gateway™ vectors, and collected as minipools of products for each target. Minipools were sequenced using the Sanger method. Eight individual colonies per minipool were isolated and repooled into eight bar-coded libraries containing one individual clone per targeted gene. Barcoded libraries were sequenced using Illumina and Roche/454 platforms. Minipool and deconvolved single-clone sequences were trimmed for vector and barcode sequences, filtered for quality, and aligned to the WS190/CE6 genome sequence. Alignments that extended beyond the CDSspecific primer were annotated and consolidated with other 3'UTRome datasets in AceView and 3'UTRome databases (see Supplementary Materials and Methods for details).

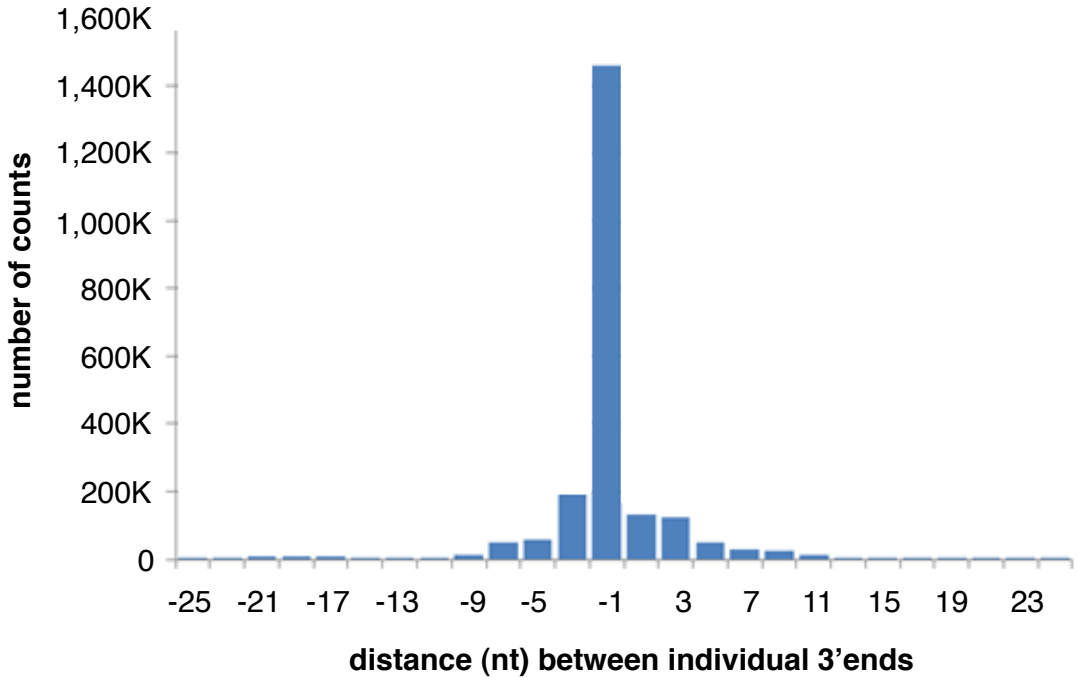


Figure 4.10 Distance between individual 3' ends and the representative polyA addition site for a cluster.
Frequency distribution of distance (in nucleotides) between the representative polyA site in a cluster and all other polyA sequence tags in the same cluster. Data are cumulative for all polyA clusters in the 3'UTRome. 86% of individual polyA tags fall within 4nt of the representative polyA site.

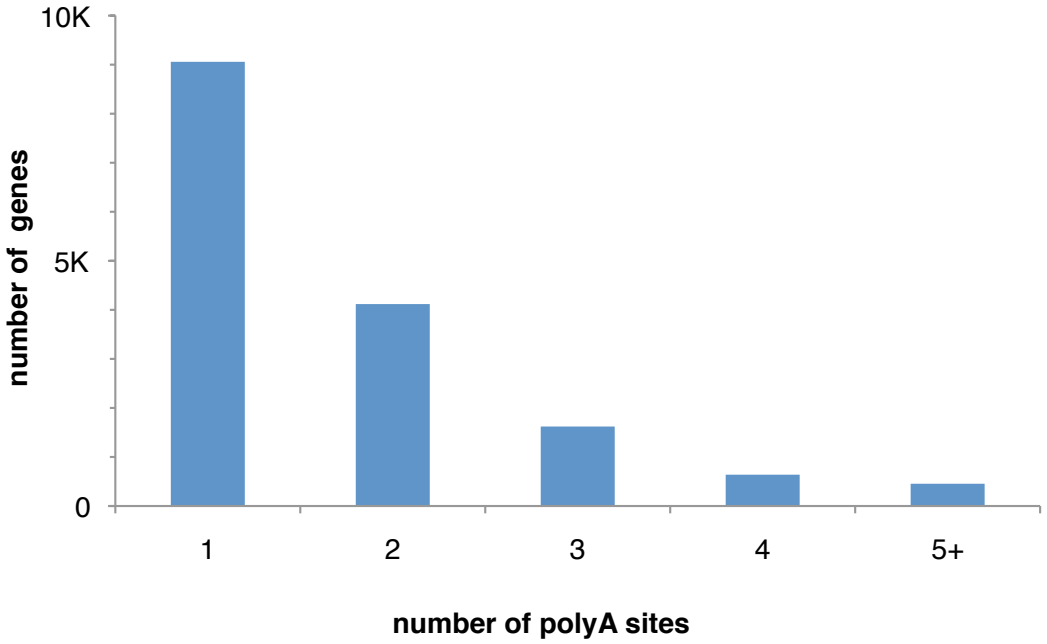


Figure 4.11 Number of polyA sites per gene.
The frequency distribution of distinct representative polyA sites per gene in the 3'UTRome. Around 40% of all genes with an annotated 3'UTR contain more than one alternative polyA site. Among genes with a large number of alternative 3'UTR isoforms are those encoding the small GTPase RAB-11.1 (6 isoforms), the LIN-61 paralog MBTR-1 (7 isoforms), and the RNA helicase VBH-1 (8 isoforms).

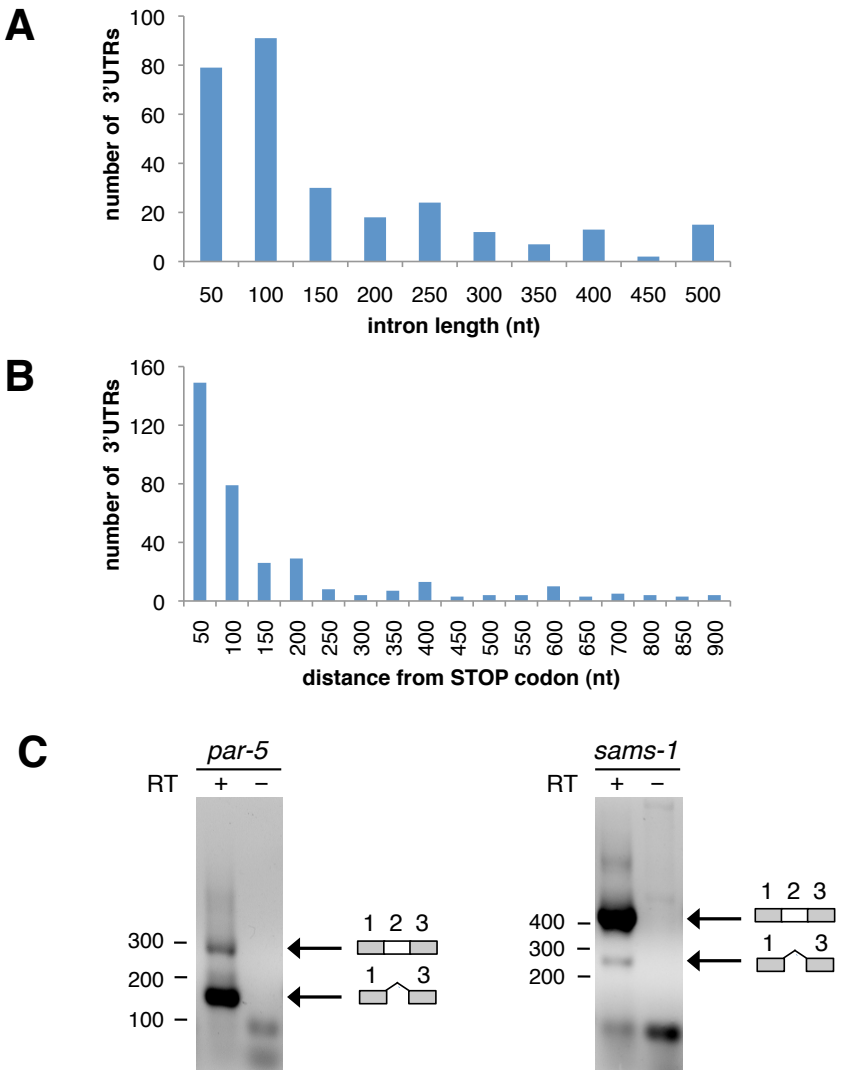


Figure 4.12 Introns in 3'UTR regions.
363 intron-containing 3'UTRs for 192 unique genes were used in this analysis. A) Length distribution (in nucleotides) of introns in 3'UTRs. B) Length distribution of the distance from the STOP codon to the intron start position. In both A and B, intron length is shown in 50 nt bins for simplification. C) Examples of facultative introns. Shown are 3'RACE products from *par-5* and *sams-1* 3'UTRs using mixed-stage total RNA and gene-specific primer pairs flanking the intron (regions 1 and 3), with (+) or without (-) inclusion of reverse transcriptase (RT) in the reaction. Agarose gel electrophoresis lanes with RT each produce two products consistent in size with the retention (top band) or excision (lower band) of region 2. Small bands below 100 nt represent unamplified primers and primer dimers (see Supplementary Online Materials and Methods for details). We observe that in some of these 3'UTRs, putative binding sites for miRNAs or ALG-1 (200) are contained within an intronic sequence.

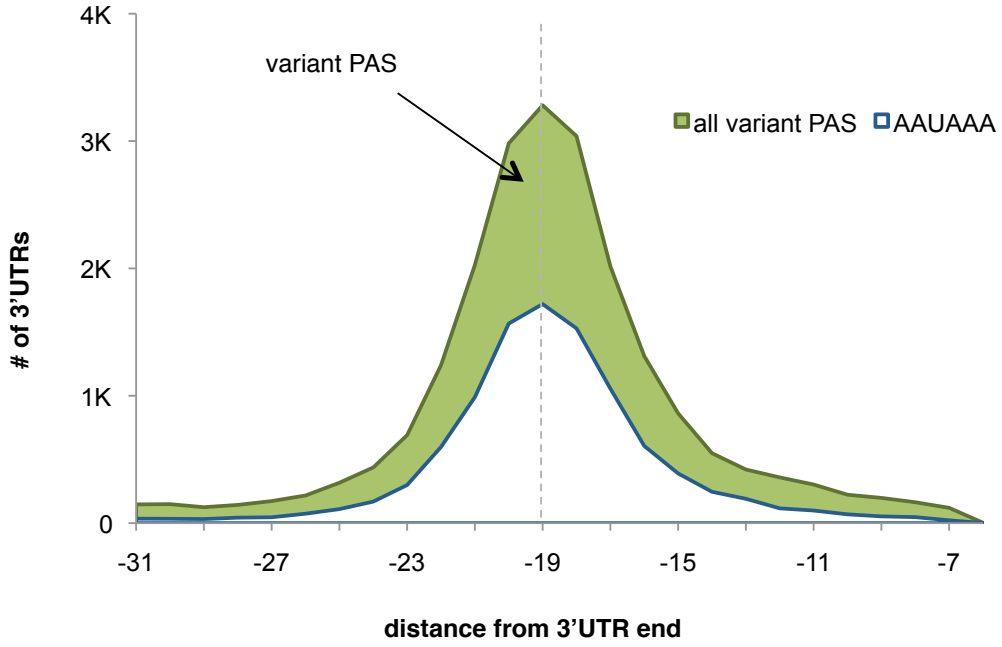


Figure 4.13 Distribution of the canonical AAUAAA and variant PAS elements relative to the cleavage and polyA addition site.
Start position for all PAS motifs (green line), AAUAAA (blue line), and variant PAS (green shading) peak at 19 nt upstream of the polyA addition site. See Supplementary Materials and Methods for details on the identification of PAS motifs and assignment of the most likely PAS motif for each 3'UTR.

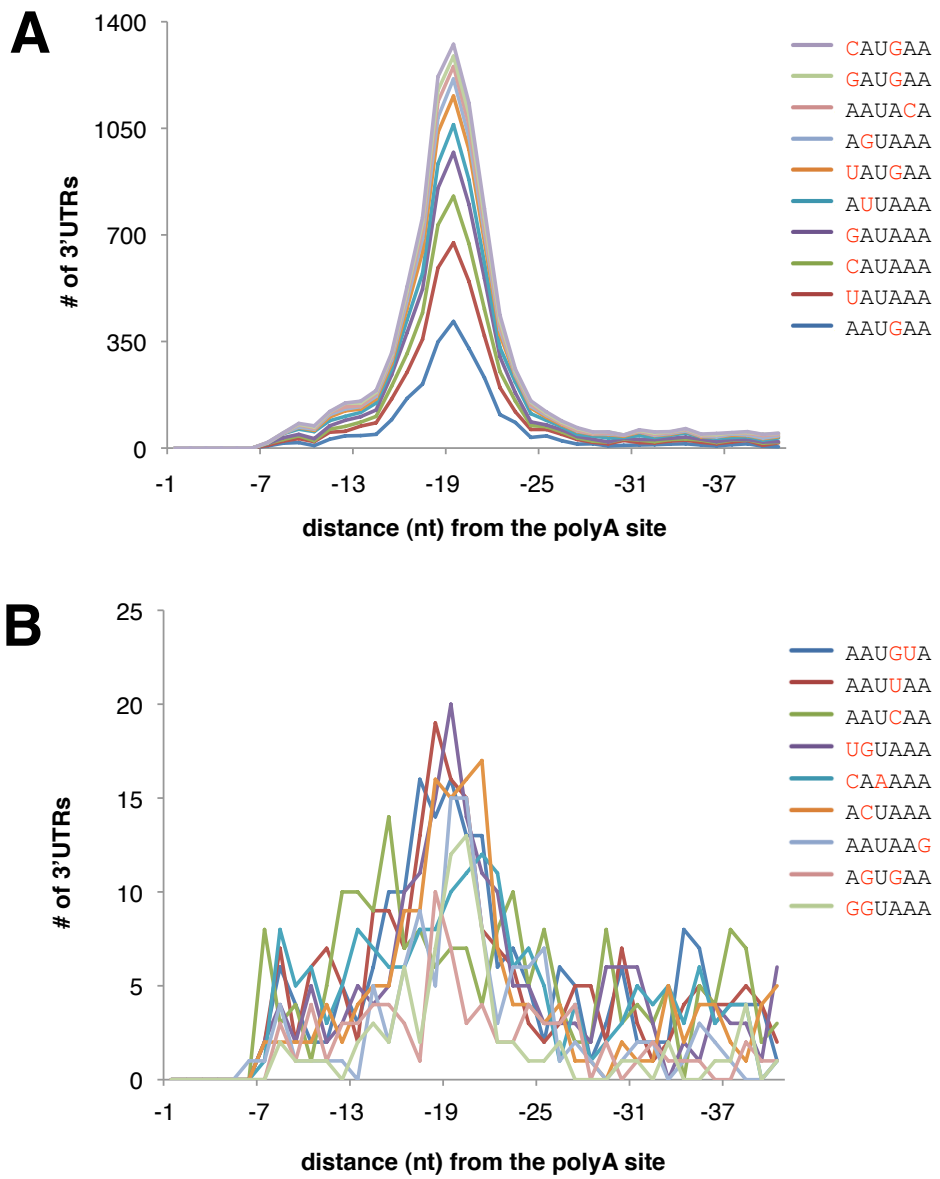


Figure 4.14 Distribution of variant PAS elements relative to the cleavage and polyA addition site.
In an unbiased search of all possible hexamers in the regions upstream of polyA sites in the 3'UTRome, the most common variant PAS hexamers show an enrichment that peaks at 19-20 nucleotides upstream of the polyA site. Using this as a guide, the most likely PAS motif for each polyA site was assigned using an ordered list of motifs according the frequency of each motif in this region (see Supplementary Materials and Methods for details). The distribution of the most common motif, the canonical AAUAAA, which peaks at position -19, is not shown in this figure. A) Ten of the most common variant PAS motifs (each assigned to ≥1% of all polyA sites). The most common PAS variants contain a U in the third position and an A in the sixth position. B) Nine of the least common variant PAS motifs (each assigned to ≤1% of all polyA sites). Total counts for each motif are given in Table 4.5.

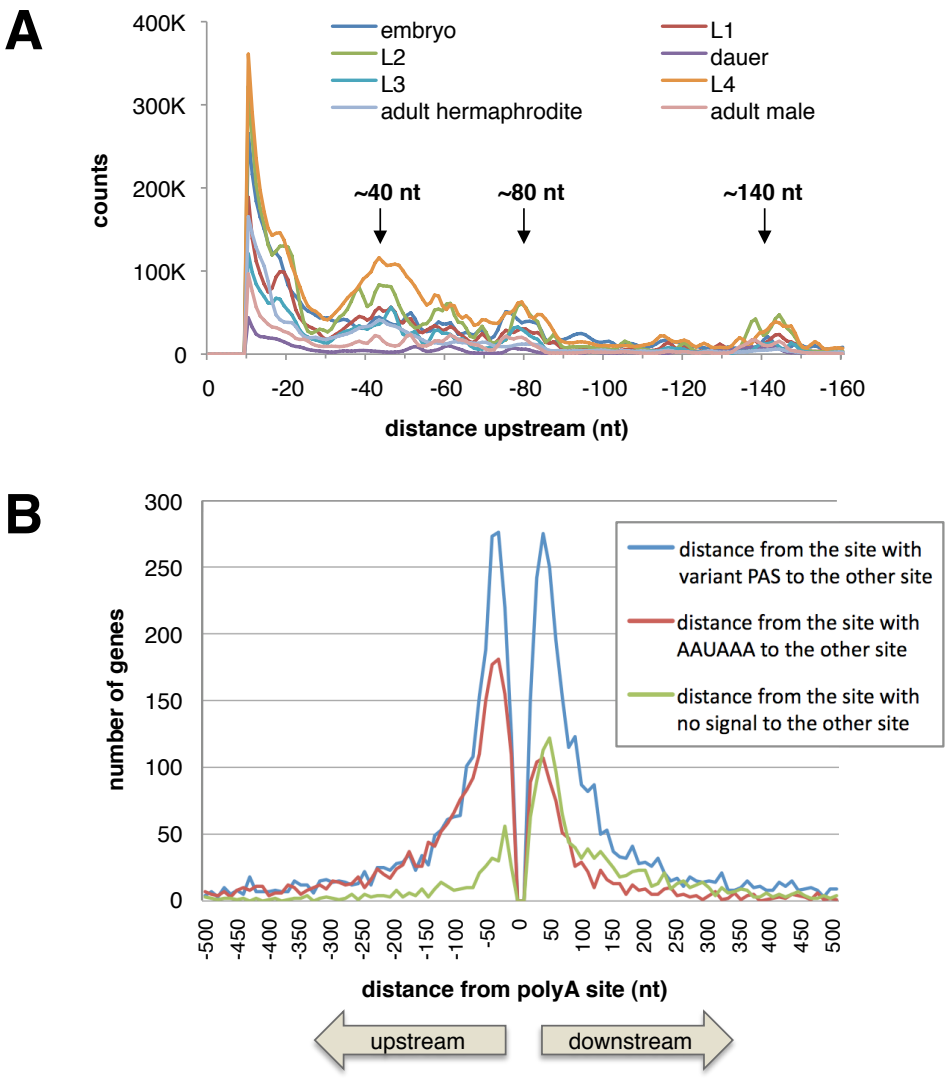


Figure 4.15 Relationship between alternative polyA addition site for the same transcript.

A) The autocorrelation of polyA addition sites, pooled by stage, showing the average support count at each position relative to the most highly supported polyA site (aligned at 0 nt). The data show a main peak (arrow) ~40-45 bases upstream of the dominant polyA site. B) The distance between adjacent polyA sites peaks at ±45nt. PolyA addition sites with the canonical AAUAAA PAS motif (red) show a propensity to have a neighboring polyA site upstream; conversely, sites with no detectable PAS (green) tend to have a neighboring site downstream. Sites with a variant PAS (blue) are equally likely to have a neighboring site upstream or downstream.

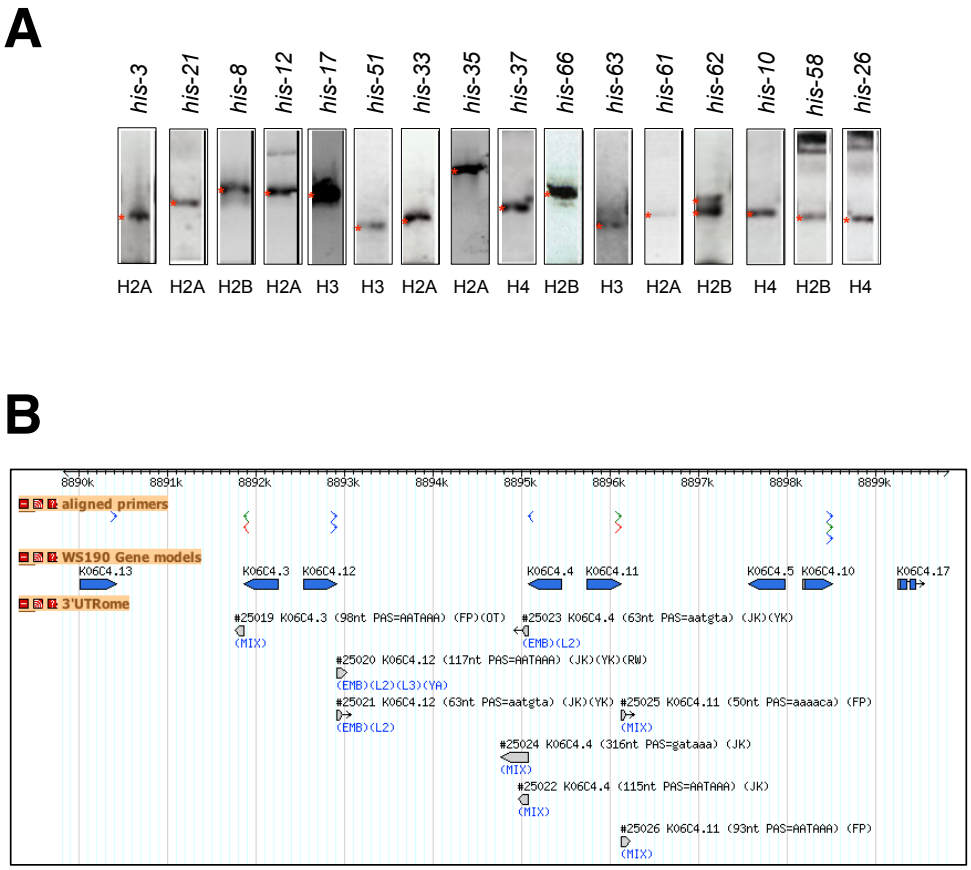


Figure 4.16 Polyadenylated 3'UTRs for histone genes.
A) The electrophoretic analysis on 2% agarose E-Gels of selected 3'RACE clones corresponding to 3'UTRs of histone genes obtained with the 3'RACE pipeline. PCR amplicons (red asterisks) correspond to unique or multiple 3'UTR isoforms. B) Histone gene cluster on chromosome V. Several histone genes with corresponding 3'UTRs detected in multiple developmental stages are shown. See Table 4.6 for the comprehensive list of histone 3'UTRs and PAS usage.
Combined with the observation that depletion of the SLBP homolog CDL-1 by RNAi severely depletes histone protein but not mRNA levels (198), our data lend support to the hypothesis that replication-dependent histone transcripts in *C. elegans* are first cleaved and polyadenylated using a PAS-directed mechanism, and are later post-processed to their final stem-loop form and regulated at the translational level by factors including CDL-1.

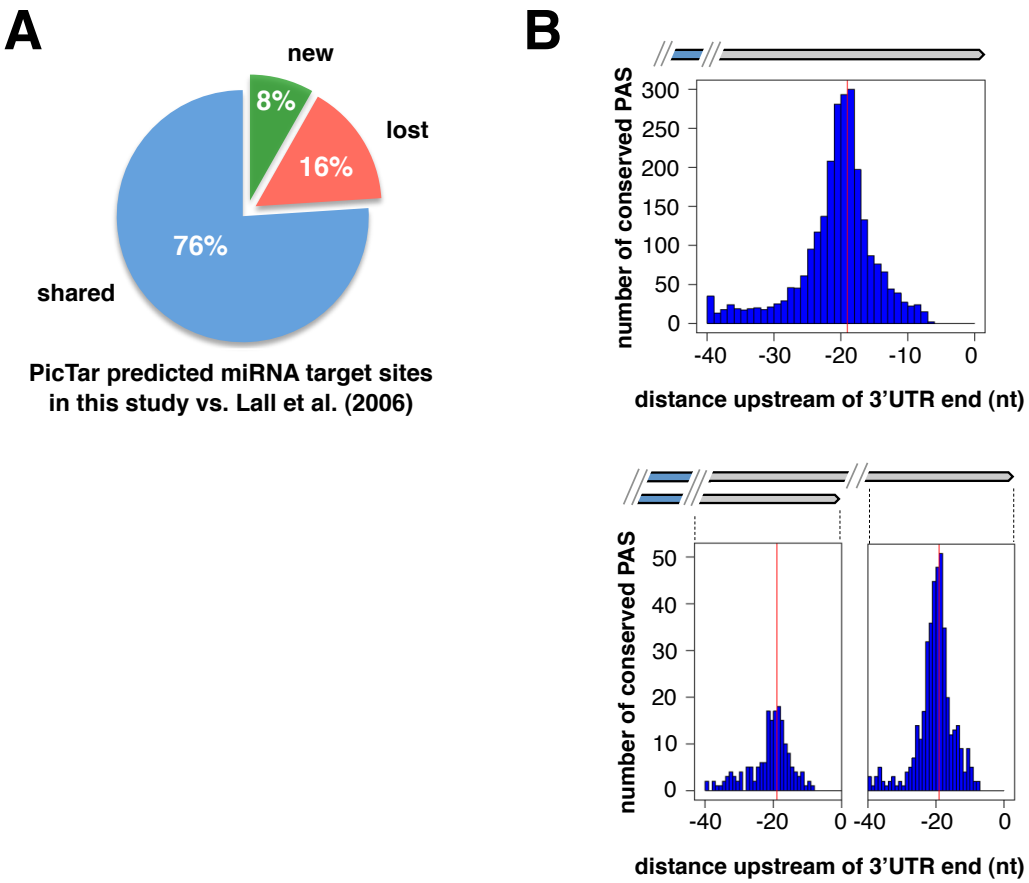


Figure 4.17 PicTar target predictions and PAS conservation in UTRome 3'UTRs.

A) Differences in PicTar predicted miRNA target sites within sequences spanned by the 3'UTRome, from this study in comparison with our previous predictions for *C. elegans* (199), as a percentage of the total number of predictions from both studies. See also Table 4.7. B) Distribution of conserved PAS motifs within 40 nt upstream of 3'UTR ends in three-way alignments between *C. elegans*, *C. briggsae*, and *C. remanei*, for (top) genes with one isoform (n=2,573 3'UTRs) or (bottom) exactly two isoforms (short, n=173; long, n=419). Red lines indicate the peak at -19 nt from the 3'UTR polyA addition site. See Supplementary Materials and Methods for additional details.

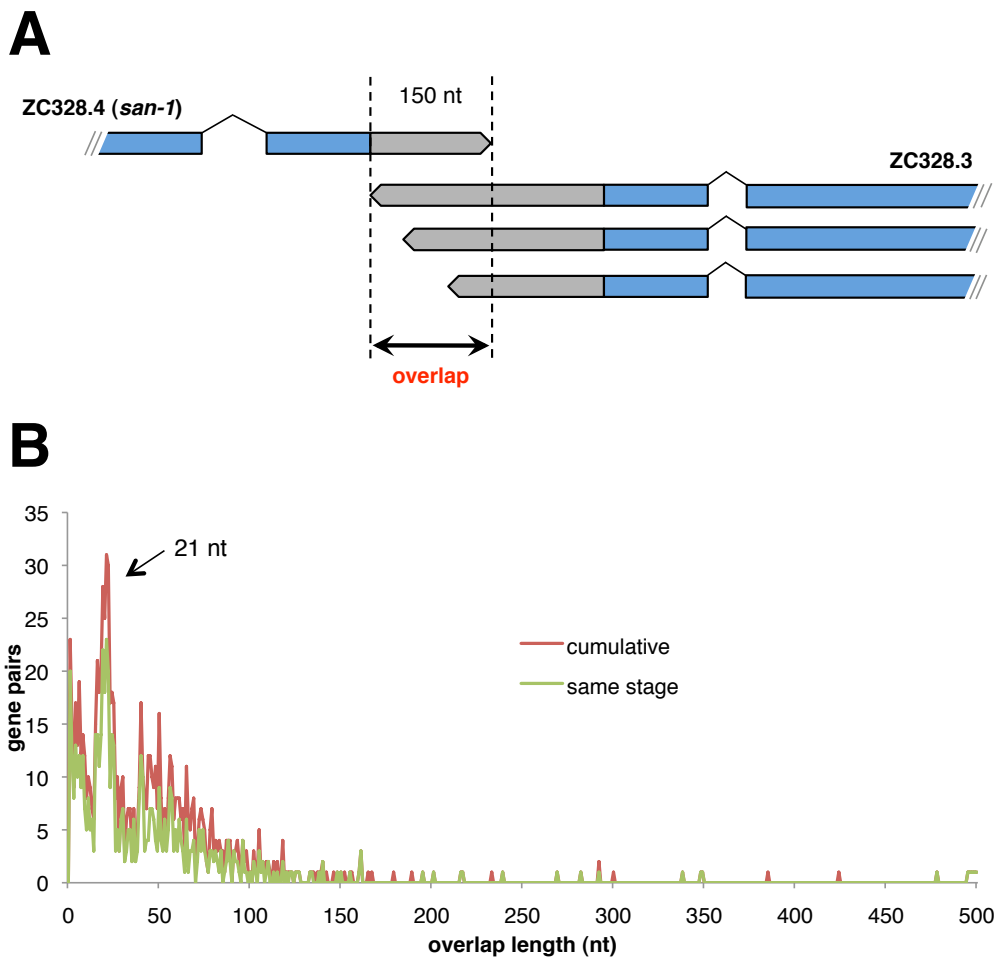


Figure 4.18 3'UTRs on opposite strands sometimes overlap.
The 3'UTRome contains 1,876 convergently transcribed neighboring genes with overlapping regions that extend from the distal end of each putative transcript into the 3'UTR or CDS of the neighboring gene. For 1,240 of these genes, overlapping 3'UTR isoforms are co-expressed during at least one developmental stage. If both genes are transcribed simultaneously in the same cell, their 3'UTRs could potentially pair as dsRNA and trigger the production of endogenous siRNAs (endo-siRNAs) (201), which could down-regulate their mRNA levels. A) Example of a 3'UTR overlap between the gene encoding mitotic spindle checkpoint protein ZC328.4 (*san-1*) and the uncharacterized gene ZC328.3. B) Length distribution (nt) of overlapping 3' end annotations for gene pairs on opposite strands, for cumulative overlapping pairs (red, n=938 pairs) or pairs detected in the same developmental stage (green, n=620 pairs). Overlapping pairs involve ~10% of genes in the 3'UTRome. Overlaps range from 1 to 495 nt, with an average overlap length of ~44 nt and median overlap length of ~28 nt. The peak in the overlap distribution at ~21 nt suggests that longer overlaps generally may be disfavored to limit recruitment of cellular machinery that could lead to endo-siRNA production (201).

Table 4.1 Sequence data in the 3'UTRome.

datasets	platform	total sequences	mapped sequences	developmental stage data	distinct polyA supported
PolyA capture	454	2,532,433	2,138,657	YES	165,538
RACE clones	Sanger	7,105	5,139	No	44,807
	454	166,112	86,577	No	
	Illumina	49,958,257	9,693,792	No	
cDNA	Sanger	—	119,434	YES	57,048
RNA-Seq	Illumina	291,573,831	84,771	YES	37,220

Total number of raw and mapped sequences and the number of distinct polyA clusters supported for each data stream. Three of the datasets, polyA capture, cDNA and RNA-seq, provide developmental stage information allowing us to link distinct 3'UTR isoforms to specific developmental stages. See Figure 4.6-Figure 4.9 for details on the different pipelines.

Table 4.2 Summary of the polyA capture 454 sequencing runs.

RUN 1		embryo	L1	L2	L3	L4	adult	male
total sequences		631,599	277,370	424,818	289,673	341,573	151,172	206,624
barcode detected		565,640	265,441	422,739	272,074	336,304	150,835	202,348
usable	**yes**	560,522	262,071	417,695	269,815	332,494	146,549	201,236
	no	5,118	3,370	5,044	2,259	3,810	4,286	1,112

RUN 2		*daf-2*	*daf-7*	*daf-9*	*daf-11*
total sequences		87,880	60,335	51,781	64,931
barcode detected		76,729	53,798	50,551	53,779
usable	**yes**	76,200	53,429	50,234	53,429
	no	529	369	317	315

Roche/454 reads produced by the polyA capture in individual developmental stages, males, and dauer mutants. The sequences obtained (total sequences) were scanned for the detection of a barcode (barcode detected). Reads containing a sequence contiguous with a polyA site were classified as 'usable'.

Table 4.3 Gene and 3'UTR isoform coverage for individual datasets and overlap between datasets in the 3'UTRome using Aceview gene models.

	PolyA capture	3'RACE	cDNA	RNA-seq
PolyA capture	**11,606 (17,131)**	—	—	—
3'RACE	3,879 (4,419)	**5,929 (7,707)**	—	—
cDNA	7,845 (9,808)	3,981 (4,475)	**11,447 (16,986)**	—
RNA-seq	5,445 (5,945)	2,732 (2,878)	6,040 (6,686)	**7,442 (8,332)**
total	**15,683 (26,942)**			
specific genes	1,858 (5,453)	632 (1,955)	1,358 (4,714)	314 (549)

Diagonal cells show the total number of coding genes and distinct polyA ends (in parentheses) for each of the four independent datasets; off-diagonal cells show intersections between each pair of datasets. The last row shows the total number of coding genes and distinct polyA ends that are specific to each individual dataset.

Table 4.4 Subset of 3'UTRome matching WS190 gene models.

	PolyA capture	3'RACE	cDNA	RNA-seq
PolyA capture	**11,007 (16,151)**	—	—	—
3'RACE	3,878 (4,399)	**5,919 (7,641)**	—	—
cDNA	7,853 (9,724)	3,994 (4,469)	**11,387 (16,710)**	—
RNA-seq	5,394 (5,841)	2,743 (2,875)	6,070 (6,652)	**7,322 (8,130)**
total	**14,986 (25,650)**			
specific genes	1,382 (4,658)	635 (1,915)	1,300 (4,547)	253 (473)

The subset of data from Table 4.3 that are compatible with WormBase WS190 gene models. See Table 4.3 legend for additional details.

Table 4.5 Identification of putative PAS elements.

name	3'UTRs	frequency (%)
AAUAAA	10,797	38.9
no PAS	3,658	13.2
AAUGAA	2,576	9.3
UAUAAA	1,731	6.2
CAUAAA	1,021	3.7
GAUAAA	974	3.5
UAUGAA	759	2.7
AUUAAA	746	2.7
AAAAAA	660	2.4
UUUAAA	487	1.8
AGUAAA	416	1.5
AAUACA	387	1.4
AAUAUA	353	1.3
GAUGAA	313	1.1
AAUAAU	311	1.1
CAUGAA	310	1.1
AAAUAA	307	1.1
UGUAAA	302	1.1
UCUAAA	231	0.8
AAUGUA	229	0.8
AAUUAA	176	0.6
ACUAAA	174	0.6
AAGAAA	168	0.6
CAAAAA	167	0.6
GAAAAA	146	0.5
AACAAA	115	0.4
AAUAAG	93	0.3
GGUAAA	92	0.3
AGUGAA	55	0.2
AAACAA	35	0.1

An unbiased search for over-represented hexamers in the last 50 nt of 3'UTRs in the 3'UTRome identified a handful of sequences whose start positions all peaked at around 19 nt upstream of the polyA cleavage site. Using these results as a guide, we searched all 3'UTRs recursively for the most likely PAS site utilized by each 3'UTR (see Supplementary Materials and Methods for details). The most common motif, the "canonical" PAS element AAUAAA, is observed in 39% of 3'UTRs; the other elements consist of variations of this motif differing by one or two nucleotides. This apparent diversity of PAS motifs suggests that the recognition of PAS sites in worms is more flexible than higher eukaryotes, where mutation in any position of the canonical AAUAAA element disrupts the 3'end processing of mRNAs (213), and may perhaps be more akin to the 3' end processing mechanism of yeast, where presence of an AU rich region is sufficient to allow docking of the processing machinery (214).

Table 4.6 Cumulative list of polyadenylated 3'UTRs detected in histone genes.

name	CDS	isoforms	3'UTR length (PAS)
his-2	T10C6.13	3	48 (no signal), 127 (AAUAAA), 365 (no signal)
his-3	T10C6.12	1	97 (AAUAAA)
his-4	T10C6.11	2	14 (no signal), 112 (AAUAAA)
his-6	F45F2.13	1	128 (AAUAAA)
his-8	F45F2.12	1	66 (no signal)
his-9	ZK131.3	2	120 (AAUAAA), 180 (AAUAAA)
his-10	ZK131.4	1	114 (AAUAAA)
his-11	ZK131.5	1	108 (GAUAAA)
his-12	ZK131.6	1	97 (AAUAAA)
his-13	ZK131.7	1	120 (AAUAAA)
his-14	ZK131.8	1	114 (AAUAAA)
his-15	ZK131.9	1	111 (GAUAAA)
his-16	ZK131.10	2	58 (no signal), 119 (AAUAAA)
his-19	K06C4.11	2	50 (AAAACA), 93 (AAUAAA)
his-20	K06C4.4	3	63 (AAUGUA), 115(AAUAAA), 316 (GAUAAA)
his-21	K06C4.3	1	98 (AAUAAA)
his-22	K06C4.12	2	63 (AAUGUA), 117 (AAUAAA)
his-24	M163.3	1	236 (AAUAAA)
his-25	ZK131.2	3	98 (UGUAAA), 124 (AAUAAA), 151 (GAUAAA)
his-26	ZK131.1	1	114 (AAUAAA)
his-27	K06C4.13	1	222(AAUAAA)
his-28	K06C4.2	2	108 (AAUAAA), 219 (GAUGAA)
his-32	F17E9.10	2	115 (AAUAAA), 146 (AAUAAA)
his-34	F17E9.9	1	59 (AAUAAA)
his-35	C50F4.13	1	116 (AAUAAA)
his-36	C50F4.6	3	92 (AAUAAA), 100 (AAUAAA), 622 (AAUAAA)
his-37	C50F4.7	1	88 (AAUAAA)
his-40	NULL	1	128 (AAUAAA)
his-41	C50F4.5	3	92 (AAUAAA), 100 (AAUAAA), 622 (AAUAAA)
his-42	F08G2.3	1	276(AAUAAA)
his-43	F08G2.2	1	97 (AAUAAA)
his-44	F08G2.1	1	111 (GAUAAA)
his-45	B0035.10	1	116 (AAUGAA)
his-46	B0035.9	4	29 (no signal), 67 (no signal), 114 (AAUAAA), 155 (AAUACA)
his-47	B0035.7	2	115 (AAUAAA),171 (UAUAAA)
his-48	B0035.8	2	103 (AAUAAA), 115 (AAUAAA)
his-49	F07B7.5	1	120 (AAUAAA)
his-50	F07B7.9	2	108 (AAUAAA), 219 (GAUGAA)
his-51	F07B7.10	1	93 (AAUAAA)
his-52	F07B7.4	2	63 (AAUGUA), 117 (AAUAAA)
his-53	F07B7.3	2	50 (AAAACA), 93 (AAUAAA)
his-54	F07B7.11	3	63 (AAUGUA), 115 (AAUAAA), 316 (GAUAAA)
his-56	F54E12.3	3	29 (no signal), 67 (no signal), 114 (AAUAAA)
his-57	F54E12.5	1	104 (AAUAAA)
his-58	F54E12.4	1	103 (AAUGAA)
his-59	F55G1.2	1	295 (AAUGAA)
his-60	F55G1.11	2	66 (no signal), 120 (AAUAAA)
his-61	F55G1.10	1	98 (AAUAAA)
his-62	F55G1.3	2	31 (no signal), 107 (AAUAAA)
his-63	F22B3.2	1	116 (AAUAAA)
his-66	H02I12.6	1	107 (AAUAAA)
his-68	T23D8.6	2	15 (AAUAAA), 100 (AAUAAA)
his-69	E03A3.3	1	90 (GAUAAA)
his-70	E03A3.4	1	106 (AAUAAA)
his-71	F45E1.6	1	163 (AAUAAA)
his-72	Y49E10.6	2	104 (AAUAAA), 213 (AAUAAA)
his-74	W05B10.1	1	162 (AAUAAA)

Summary of 3'UTR isoforms detected in histone genes, showing the putative PAS element for each representative 3'UTR. Nucleotides that deviate from the canonical PAS motif are highlighted in red.

Table 4.7 Summary statistics for PicTar miRNA target predictions and other conserved sequence blocks in genomic regions spanned by the 3'UTRome compendium.

A	# of 3'UTR isoforms	26,942	
B	# of unique 3'UTR regions	15,685	
C	average 3'UTR length	250 nt	
D	total 3'UTRome length	3,898,952 nt	
E	per nucleotide conservation rate of 3'UTR (3 species)	0.3	
F	probability of a conserved seed being functional	3 species	5 species
		0.56 ±0.01	0.64 ±0.03
G	# of unique conserved seeds identified	3 species	5 species
		5,673	1,744
H	# of unique miRNAs used for analysis (# of families)	183 (124)	
I	probability of a conserved miRNA seed site occurring inside an ALG-1 site	3 species	5 species
		0.75	0.76
J	probability of a randomly positioned 6-mer in a 3'UTR occurring inside an ALG-1 site	3 species	5 species
		0.43	0.45
K	# of conserved blocks not explained by predicted miRNA seeds or conserved PAS (5 species)	4,758	
L	# of 3'UTRs with at least one conserved block (5 species)	2,887	
M	probability of a conserved (randomly shuffled) sequence block of the same length inside an ALG-1 site	0.54 (0.48)	
M	fraction of Lall et al. 3'UTR:miRNA interactions recovered	0.83	
O	# of Lall et al. 3'UTR:miRNA interactions lost	1,111	
P	# of unique new interactions vs. Lall et al. miRNAs/3'UTRs	580	

A) Total number of 3'UTRs used for miRNA target predictions. **B)** Number of unique 3'UTR regions, obtained by merging 3'UTRs with overlapping genomic coordinates. **C)** Average length of all unique 3'UTRs. **D)** The unique 3'UTRome comprises ~4M nucleotides. **E)** 30% percent of nucleotides in *C. elegans* 3'UTR are conserved in *C. remanei* and *C. briggsae*. Nucleotides in CDS, 5'UTR or intergenic regions were not considered in this analysis. **F)** Probability of a conserved miRNA seed being functional based on alignments of three or five species, obtained by creating artificial miRNAs resembling the original miRNAs (212) and comparing the number of target sites for the artificial miRNAs with the "real" target sites. **G)** Number of unique conserved miRNA seeds in the genome of three or five species. **H)** In total, 183 miRNAs were used. They comprise 174 miRBase (database release 14) miRNAs and 9 novel miRNAs determined by miRDeep2 (211), grouped in 124 miRNA families. **I)** The probability of a conserved miRNA seed within an ALG-1 binding site (200) in three or five species, calculated as the ratio of all miRNA target sites located in an ALG-1 binding site when considering only 3'UTRs that have an ALG-1 site and at least one miRNA target site. **J)** Probability of a shuffled seed sites (randomly positioned with the same 3'UTR) occurring within an ALG-1 binding site for three or five species. The probability is 30% less for shuffled sites than for the original miRNA seed position, signifying that miRNA seeds located in ALG-1 sites are indeed accurate signals. **K)** Number of conserved blocks, defined as at least 6 nt long and present in five species, that cannot be explained by a conserved predicted miRNA target seed site or a conserved PAS. **L)** Number of 3'UTR regions that contain at least one of such conserved blocks. **M)** Probability of a conserved block occurring within an ALG-1 binding site vs. randomly positioned blocks of analyses in K-M, regions overlapping a CDS in an alternative transcript

were excluded. **N,O,P)** For the same miRNAs and 3'UTR regions, 83% of previously predicted miRNA target sites from Lall et al. (199) are identical with predictions using the empirically defined 3'UTRs in the 3'UTRome; 1,111 miRNA target sites are exclusively found in Lall et al., and 580 sites are newly predicted. Three species alignments always included *C. elegans, C. remanei*, and *C. briggsae*. Five species alignments also included *C. brenneri* and *C. japonica.* See Supplementary Materials and Methods for additional details.

Table 4.8 Number of genes present in multiple developmental stages but with stage-specific 3'UTR isoforms.

stage	genes	isoforms
embryo	966	1,320
L1	325	353
L2	252	268
dauer	264	304
L3	131	134
L4	150	157
adult	84	88
male	374	447
total	2,049	3,071

We have scanned the 3'UTRome for genes expressed in 1) at least two developmental stages, 2) with at least two 3'UTR isoforms, and 3) where one of these isoforms was stage-specific. The results shown here were used for the analysis described in Figure 4.4B.

Table 4.9 Number of genes with two 3'UTR isoforms detected in the staged polyA capture dataset.

stage	long 3'UTR more abundant	short 3'UTR more abundant
embryo	315	169
L1	80	58
L2	33	37
dauer	184	104
L3	59	34
L4	27	39
adult	80	45
male	94	97
total	915	610
total genes	1,960	

A subset of annotated genes from the polyA capture dataset with two 3'UTR isoforms used for the analyses in Figure 4.4. A 3'UTR isoform is defined as abundant if: 1) the total number of counts across all stages is larger than 5, and 2) if it is supported by at least twice the number of counts than the other 3'UTR isoform (see Supplementary Materials and Methods for details).

Table 4.10 3'UTR clones available in the 3'UTRome library.

	minipools	deconvolved library
96-well plates	75	39
unique genes	7,105	3,750
unique isoforms	—	5,774

The 3'RACE approach produced sequence-validated 3'UTR clones that are available to the community to study 3'UTR biology. The UTR library collection will be updated on an ongoing basis and will expand to contain minipools and unique 3'UTR isoforms for all *C. elegans* 3'UTRs for protein-coding transcripts. See the 3'UTR data repository http://www.utrome.org for clone availability.

Chapter 5

Pervasive and dynamic protein binding sites of the mRNA transcriptome in *Saccharomyces cerevisiae*[4]

5.1 Abstract

Protein-RNA interactions are integral components of nearly every aspect of biology including regulation of gene expression, assembly of cellular architectures, and pathogenesis of human diseases. However, studies in the past few decades have only uncovered a small fraction of the vast landscape of the protein-RNA interactome in any organism, and even less is known about the dynamics of protein-RNA interactions under changing developmental and environmental conditions. Here, we describe the gPAR-CLIP (global photoactivatable-ribonucleoside-enhanced crosslinking and immunopurification) approach for capturing regions of the untranslated polyadenylated transcriptome bound by RNA-binding proteins (RBPs) in budding yeast. We report over 13,000 RBP crosslinking sites in untranslated regions (UTR) covering 72% of protein-coding transcripts encoded in the genome, confirming 3' UTRs as major sites for RBP interaction. Comparative genomic analyses reveal that RBP crosslinking sites are highly conserved, and RNA folding predictions indicate that secondary structural elements are constrained by protein binding and may serve as

[4] Originally published in *Genome Biology* (2013;14(2):R13) with authors listed as Mallory Freeberg*, Ting Han*, James J. Moresco, Andy Kong, Yu-Cheng Yang, Zhi John Lu, John R. Yates III, and John K. Kim (* denotes equal contribution).

generalizable modes of RNA recognition. Finally, 38% of 3' UTR crosslinking sites show changes in RBP occupancy upon glucose or nitrogen deprivation, with major impacts on metabolic pathways as well as mitochondrial and ribosomal gene expression.Our study offers an unprecedented view of the pervasiveness and dynamics of protein-RNA interactions *in vivo*.

5.2 Introduction

A diverse and expanding repertoire of RNA-binding proteins (RBPs) ensures faithful expression and function of substrate mRNAs (54-56). Many RNAs are organized by RBPs and other protein co-factors into higher-order ribonucleoprotein (RNP) assemblies that fulfill critical functions in storage, transport, inheritance, and degradation of RNA (57, 58). For example, over 70% of mRNAs in *Drosophila* embryos are localized to distinct organelles, compartments, and membrane interfaces, providing a means for directing local translation and regulating cellular architectures and functions (3). RNA and RBPs can also reversibly aggregate into granules to allow RNA storage and decay in response to stimuli (61, 62). These and many other processes are driven by large, complex networks of protein-RNA interactions that provide specificity in gene regulation and fidelity in RNP assembly. Despite important insights regarding the necessity of RNA regulation for cellular functions, the RBP-RNA interactome and its response to changing cellular conditions have yet to be fully elucidated.

Here, we adapt the PAR-CLIP technique to map all RBP binding sites across the yeast non-translating mRNAs in different environmental conditions, a method we call global PAR-CLIP (gPAR-CLIP). The comprehensive identification of RBP-RNA crosslinked sites visualized by gPAR-CLIP allows us to derive general properties of RBP-RNA interactions *in vivo*. Additionally, we compared RBP-RNA crosslinked sites in rapidly proliferating versus stress-treated cells and observed large-scale changes in RBP-RNA interactions, providing a starting point for dissecting the network of post-transcriptional gene regulatory mechanisms underlying stress response.

5.3 Results

5.3.1 gPAR-CLIP identifies transcriptome-wide RBP crosslinking sites

To construct a global map of RBP binding sites on the transcriptome *in vivo*, we combined PAR-CLIP with high-throughput sequencing (Figure 5.1A; see Methods). Briefly, we metabolically incorporated the photoactivatable-nucleobase analog 4-thiouracil (4sU) in growing yeast and used UV irradiation to crosslink 4sU to juxtaposed proteins, “freezing” protein-RNA interactions *in vivo*. Next, we implemented three biochemical strategies to capture RNA regions bound by the proteome: (i) sucrose gradient centrifugation to reduce ribosome abundance; (ii) oligo(dT) selection to deplete abundant structural non-coding RNAs (e.g. rRNAs); and (iii) chemical biotinylation of proteins. We exploited the high-affinity streptavidin-biotin interaction to purify all biotin-protein-RNA complexes with high efficiency and stringency. After trimming unbound RNA, RBP-protected

fragments were ligated to linkers, converted to cDNAs, and subjected to Illumina high-throughput sequencing (see Methods). We term this global PAR-CLIP method “gPAR-CLIP”. There are two caveats associated with our gPAR-CLIP protocol. First, during crosslinking (~5 minutes), the cells were in nutrient-free buffer and incubated on ice, which could trigger changes in RBP binding. Second, we limited our analysis to mRNAs from the top of the sucrose gradient, which mostly consists of non-translating mRNAs, so our conclusions apply to non-translating mRNAs.

To define the dynamic landscape of RBP-RNA interactions, we constructed duplicate gPAR-CLIP and mRNA-seq libraries, including incorporation of 4sU, for the wild-type *S. cerevisiae* strain cultured in complete media or subjected to glucose or nitrogen starvation for two hours. An average of 10 million reads were sequenced from each gPAR-CLIP library. Of the 72% reads that mapped uniquely to the genome, over 70% contained 1 or 2 T-to-C conversion events, the signature substitution induced by 4sU crosslinking (Table 2.1). From overlapping reads, we derived clusters representing RNA regions crosslinked to proteins (Figure 5.8 and Figure 5.9; see Methods). Crosslinking site read coverage was normalized to mRNA expression levels calculated as reads per million mapped reads per kilobase of transcript (RPKM). Because our approach captures protein-RNA interactions for potentially all RBPs, we cannot rule out the possibility that some clusters are located proximally to true RBP-binding sites (215); therefore, we refer to our gPAR-CLIP read clusters as “crosslinking sites”. We empirically assigned a false discovery rate (FDR) to each

crosslinking site by deriving clusters from mRNA-seq reads with 1 or 2 T-to-C mismatches representing sequencing error and comparing the T-to-C conversion rate of these clusters to those derived from gPAR-CLIP reads (see Methods). Using a 1% FDR threshold, we reproducibly identified 80,883 crosslinking sites that are, on average, 23 nt long (Figure 5.1B, C): 65,992 in protein-coding sequences (CDSs), 4,508 in 5' untranslated regions (UTRs), 8,525 in 3' UTRs, and 818 in introns (Figure 5.9). 6,228 (93%) of 6,717 annotated protein-coding transcripts have at least one crosslinking site. Because CDS crosslinking sites exhibited 3-nt periodicity, a hallmark of ribosome binding (Figure 5.10), we separately analyzed CDS, 5' UTR, and 3' UTR crosslinking sites. We observed high correlation between gPAR-CLIP read coverage of 5' UTRs and CDSs (Pearson correlation coefficient, R^2=0.45), supporting a prominent role for 5' UTRs in translational regulation (Figure 5.1D). However, gPAR-CLIP read coverage of 3' UTRs correlated poorly with both 5' UTRs (R^2=0.20) and CDS (R^2=0.21), suggesting a greater role in post-transcriptional regulation. As expected, correlation between total mRNA-seq read coverage across all transcript regions was approximately equal (Figure 5.1D). We also observed very high correlation of gPAR-CLIP and mRNA-seq read coverage between replicates over each genic region (Figure 5.1D), reflecting low technical variation between replicates.

5.3.2 gPAR-CLIP captures known and novel crosslinking sites

To assess the performance of gPAR-CLIP in capturing known RBP-RNA interactions, we evaluated its ability to identify binding sites of Puf3p, a Pumilio-family RBP, which we derived from conventional PAR-CLIP using a strain expressing a TAP-tagged Puf3p fusion protein. Of the 1,236 Puf3p binding sites confidently identified by PAR-CLIP, 1,008 (82%) were also captured by gPAR-CLIP (Figure 5.2A); for example, the two functionally validated Puf3p binding sites in the *COX17* 3' UTR were identified by both Puf3p PAR-CLIP and gPAR-CLIP (Figure 5.2B) (7, 216). It is possible that other Puf proteins with similar RNA recognition motifs are binding at these sites in our gPAR-CLIP libraries (217), reflecting that our protocol does not distinguish the RBPs associated with each crosslinking site. From our Puf3p PAR-CLIP library, we also identified 560 novel Puf3p mRNA targets harboring a binding site containing a Puf3p recognition motif (Figure 5.11). Given the high recovery of Puf3p sites by gPAR-CLIP, we conclude that gPAR-CLIP faithfully captures binding sites of a known RBP.

We next examined our data for general RBP-RNA interaction signatures related to mRNA maturation and translational regulation. gPAR-CLIP read coverage of 5' UTRs peaked within 75 nt downstream of annotated transcription start sites but was reduced when yeast were grown in media lacking glucose or nitrogen (Figure 5.2C). This coverage likely reflects RBP-RNA interactions involved in translation initiation, and the decrease in coverage is consistent with decreased translation initiation that occurs during cellular stress (218-220). gPAR-CLIP also effectively captured the spliceosome binding pattern by

identifying intronic RBP crosslinking sites clustering 3' of the lariat branch point (BP) bound by the U2 snRNP (Figure 5.2D). These crosslinking sites contain the canonical BP-binding protein recognition sequence UACUAAC (221, 222). Consistent with stress-induced transcriptional repression of ribosomal subunits (223), which account for 18% of all protein-coding genes with introns, gPAR-CLIP read coverage at the lariat BPs of ribosome-encoding mRNA introns decreased upon glucose and nitrogen deprivation. Finally, a strong RBP crosslinking signature was identified ~20 nt upstream of the most prominent poly(A) junction site identified in each 3' UTR (Figure 5.2E), consistent with interactions with the polyadenylation complex (224). Taken together, these results indicate that gPAR-CLIP faithfully captures diverse RBP-RNA interactions along the discrete anatomy of mRNAs.

5.3.3 RBP crosslinking sites exhibit global conservation in both primary sequences and secondary structures

Compared to mRNA-seq reads, which were equally distributed among 5' and 3' UTRs and CDSs, gPAR-CLIP reads were 4-fold enriched in 3' UTRs, 2.5-fold enriched in 5' UTRs, and 4-fold depleted in CDSs compared to mRNA-seq reads (Figure 5.3A). To examine RBP binding activity at nucleotide resolution, we calculated a crosslinking score (CLS) for each T in the genome (U in the transcriptome) as the ratio of gPAR-CLIP reads containing 1 or 2 T-to-C conversion events to mRNA-seq reads to normalize for variable mRNA abundance (see Methods). 378,247 Ts (12.7% of transcriptomic Us) were assigned a CLS: high CLS values indicate high crosslinking efficiency and strong

RBP-RNA interactions; low CLS values indicate low crosslinking efficiency or weak/transient RBP-RNA interactions. Consistent with the distribution of gPAR-CLIP reads, CLS values were highest in 3' UTRs followed by 5' UTRs and CDSs (Figure 5.3B; Figure 5.12). These observations support 3' UTRs as the primary sites for RBP-RNA interactions for non-translating mRNAs. To determine if enrichment of gPAR-CLIP reads on UTRs was biased because of the U-richness of UTRs, we compared the proportion of Us in each crosslinking site to its coverage in gPAR-CLIP and observed only a weak positive correlation, which by itself cannot account for the 4-fold enrichment of gPAR-CLIP reads on UTRs (Figure 5.3C).

A previous comparative analysis of seven *Saccharomyces* genomes revealed that ~14% of evolutionarily constrained bases lie outside protein-coding regions, often located in UTRs (225). These conserved regions could represent functional elements interacting with *cis*-acting factors. We found direct evidence of RBPs crosslinking to 35% of conserved sequence blocks in UTRs as defined by phastCons, a score representing the likelihood that a base falls in a conserved element (Figure 5.3D): 405 of 1,549 5' UTR blocks (26%) and 1,036 of 2,536 3' UTR blocks (41%) completely overlap with at least one RBP crosslinking site, which is significantly higher than randomly defined control blocks (χ^2 test, $P<10^{-119}$ for 3' UTR and $P<10^{-22}$ for 5' UTR).

At the gene level, *ATG8*, a key autophagy gene, contains two major crosslinking sites that overlap with conserved sequence blocks in its 3' UTR (Figure 5.3E, top; Figure 5.12). Similarly, *TOM40*, which encodes a translocase

that mediates import of mitochondria-localized proteins into the mitochondria, contains two major 3' UTR crosslinking sites in regions with high local conservation (Figure 5.3E, bottom; Figure 5.12). To further elucidate the connection between RBP binding and conservation, we binned Ts by CLS values and observed that Ts in all 3' and 5' UTR bins, as well as the majority of CDS (78%) bins, were more conserved than randomly binned Ts, suggesting that RBP crosslinking sites are under purifying selection (Figure 5.3F; Figure 5.12).

Unexpectedly, 3' and 5' UTR nucleotides in the lowest CLS bins exhibited extremely high conservation. Since a low CLS can indicate inefficient RNA capture, and gPAR-CLIP inefficiently captures highly structured, double-stranded RNA (see Discussion), we hypothesized that low CLS/high conservation bins represent conserved, secondary structure motifs recognized by RBPs. For example, She2p binds a distinct stem-loop structure in several bud-localized mRNAs (226, 227), including *ASH1*, for which the She2p 3' UTR recognition element is weakly represented in our gPAR-CLIP dataset. To determine if Ts with low CLS values are located in RNA regions with a high degree of secondary structure, we computed the probability of each T being unpaired using RNAplfold, a local thermodynamic folding algorithm (228). We observed that Ts with low CLS values exhibited low unpaired probabilities, suggesting they are more likely to exist in double-stranded structures (Figure 5.4A; Figure 5.13). Additionally, a strong, positive correlation between unpaired probability and CLS values indicates that unpaired regions crosslink more strongly to RBPs. To probe RNA secondary structures more accurately, we extended the boundaries of each

crosslinking site to span 80 nt and calculated the most thermodynamically stable secondary structure. Consistent with the per nucleotide analysis, crosslinking sites with low CLS values formed predominantly double-stranded RNA structures (Figure 5.4B; Figure 5.13).

Secondary structures tolerate substitutions that preserve base pairing in stem regions, a characteristic known as covariance. To identify conserved and thermodynamically stable RNA secondary structures using a covariance model, the seven yeast genomes were scanned with RNAz (229, 230), and a small set of potential structural elements was identified: 843 in CDS, 25 in 5' UTRs, and 51 in 3' UTRs. Among Ts assigned a CLS, those with the lowest CLS values in CDS and 3' UTRs were preferentially located in conserved, structural elements compared to control elements (Figure 5.4C; Figure 5.13). Taken together, our per nucleotide and per crosslinking site results indicate that high conservation observed for Ts with low CLS values is driven by conserved RNA secondary structures, while Ts with high CLS values are located in exposed, single-stranded RNA regions available for sequence-specific contact with RBPs.

5.3.4 Large-scale changes of RBP crosslinking site occupancy occur upon nutrient deprivation

To explore RBP-RNA interaction dynamics under changing cellular conditions, we compared gPAR-CLIP read coverage of individual 3' UTR crosslinking sites between glucose or nitrogen starvation and log-phase growth conditions. As we selected non-translating mRNAs for gPAR-CLIP analyses, we

cannot distinguish whether the changes in binding site coverage reflect changes in RBP binding or changes in RBP distribution in the sucrose gradient (see discussion). We only examined crosslinking sites with >5 RPM in gPAR-CLIP libraries to ensure confident quantification (Figure 5.14; see Methods). The intra-replicate variation of crosslinking site read coverage was quantified as standard deviation σ=1.3-fold (Figure 5.14); therefore, we consider crosslinking sites with more than 4-fold (3σ) differences in read coverage between WT and stress conditions as "increased" or "decreased". We observed >4-fold changes in crosslinking site coverage, also referred to as "RBP occupancy", for 1,129 of 3,803 (30%) 3' UTR sites upon glucose starvation and for 535 of 3,932 (14%) 3' UTR sites upon nitrogen starvation (1,497 of 3,985 3' UTR sites in either condition, 38%) (Figure 5.5A, B). Similar distributions of changes were observed for crosslinking sites in 5' UTRs (Figure 5.14). Nineteen percent (116 of 623) of crosslinking sites that exhibited decreased RBP occupancy were affected by both conditions, while only 5% (40 of 885) of crosslinking sites that exhibited increased RBP occupancy were affected by both conditions, suggesting that RBP-RNA interaction changes are largely distinct to glucose or nitrogen deprivation (Figure 5.5C). Similar to the observation that glucose starvation induced more crosslinking site occupancy changes than nitrogen starvation, comparison of mRNA abundance revealed more changes in gene expression upon glucose than nitrogen starvation (Figure 5.5D, E). Interestingly, mRNA expression of ribosomal subunits and other known RBPs was significantly down-regulated upon glucose (Welch's *t*-test, $P<10^{-27}$) and nitrogen (Welch's *t*-test,

$P<10^{-36}$) starvation, suggesting that global suppression of post-transcriptional regulation is a general response to nutrient deprivation.

We next examined the overlap of individual genes with 3' UTR crosslinking sites affected by each stress condition (Figure 5.5F). Genes harboring 3' UTR crosslinking sites with increased RBP occupancy showed little overlap (41 genes, 6%) between the two conditions; genes harboring crosslinking sites with decreased RBP occupancy showed higher overlap (114 gene, 21%). These data suggest that, for non-translated mRNA transcriptiome, loss of RBP occupancy at crosslinking sites of a larger set of common genes is a general response to nutrient limitation while increased RBP occupancy at crosslinking sites of distinct sets of genes is a nutrient-specific response.

We determined if genes exhibiting common or distinct 3' UTR crosslinking site occupancy changes under nitrogen and glucose starvation conditions had shared biological functions or cytological localization using GO enrichment analysis (Figure 5.5G, H). When we analyzed the 356 genes with sites decreased in RBP occupancy only during glucose starvation, mitochondrion-related genes and genes associated with cellular respiration were preferentially affected (Figure 5.5G, top). Analysis of the 77 genes with sites lost only during nitrogen starvation revealed enrichment for ribosomal components and noncoding RNA processing (Figure 5.5G, middle). The 114 genes harboring 3' UTR crosslinking sites with decreased coverage under both stress conditions were enriched for fatty acid and lipid catabolism (Figure 5.5G, bottom), consistent

with the utilization of stored lipids as energy source in response to nutrient deprivation (231).

Analysis of the 400 genes harboring 3' UTR crosslinking sites with increased occupancy only upon glucose starvation were enriched for terms related to translation (Figure 5.5G, top). The 254 genes harboring sites with increased RBP occupancy only upon nitrogen starvation were enriched for metabolic processes, including glutamate metabolic processes, which are affected by nitrogen availability (Figure 5.5H, middle). Of the 41 genes harboring 3' UTR crosslinking sites with increased RBP occupancy under both nitrogen and glucose starvation conditions, 13 (32%) genes represent cellular components of ribosomes or mitochondria (Figure 5.5H, bottom), consistent with induction of global changes through translational repression and changes in energy metabolism.

In order to determine whether these observations are a result of changes in mRNA abundance, we calculated GO term enrichment of mRNAs up- or down-regulated upon glucose or nitrogen starvation and observed that down-regulated mRNAs are enriched for ribosome- and translation-related genes, while up-regulated mRNAs are enriched for genes related to mitochondrion and metabolic processes (Figure 5.15). Therefore, the GO term enrichment of genes with changes in 3' UTR site occupancy cannot be fully explained by GO term enrichment of up- or down-regulated mRNAs. Taken together, these data indicate that general nutrient limitation triggers a remodeling of the post-

transcriptional regulatory programs of metabolic pathways, while glucose- and nitrogen-specific stresses affect additional, distinct biological processes.

We further visualized changes in RBP occupancy of 3' UTR crosslinking sites relative to the changes in corresponding mRNA abundance induced by glucose starvation (Figure 5.6A; Figure 5.16). Since 3' UTR crosslinking sites with decreased RBP occupancy were enriched for mitochondrion-related genes, we examined sites on a subset of these genes encoding mitochondrial membrane components and observed that the crosslinking sites were significantly depleted of RBP occupancy compared to all 3' UTR crosslinking sites (Welch's *t*-test, $P<10^{-21}$), and the mRNAs were significantly up-regulated compared to all genes (Welch's *t*-test, $P<10^{-28}$) (Figure 5.6A, blue dots). This observation suggests that 3' UTR crosslinking sites on mRNAs encoding mitochondrial membrane components are recognized by repressive RBPs, and that upon glucose deprivation, RBP-binding is attenuated, resulting in increased mRNA levels.

Mitochondrial aldehyde dehydrogenase *ALD4* mRNA is regulated transcriptionally under stress conditions (232, 233). In our gPAR-CLIP data, the *ALD4* 3' UTR harbors four highly conserved crosslinking sites displaying 2- to 8-fold decreases in RBP occupancy despite a >7-fold increase in *ALD4* mRNA levels (Figure 5.6B; Figure 5.17). These data suggest that post-transcriptional regulation of *ALD4* in response to glucose deprivation also occurs through the release of repressive RBP binding at these 3' UTR sites. *STM1*, which encodes a ribosomal subunit-associated protein required for optimal translation under

nutrient stress (234), has two 3' UTR crosslinking sites, with one exhibiting >25-fold increased RBP occupancy upon glucose starvation (Figure 5.6C; Figure 5.17). *STM1* mRNA is conversely down-regulated >3-fold, indicating a potential regulatory role for this site involving mRNA stability and/or decay. Interestingly, *STM1* mRNA expression is also down-regulated upon nitrogen starvation despite no change in RBP occupancy of this site, pointing to non-overlapping regulatory mechanisms that contribute to *STM1* regulation in glucose and nitrogen starvation conditions.

Next we explored changes in RBP occupancy of 3' UTR crosslinking sites relative to changes in corresponding mRNA abundance upon nitrogen starvation. 3' UTR crosslinking sites on mRNAs associated with ribosome biogenesis showed significantly greater decrease in RBP occupancy compared to all 3' UTR crosslinking sites (Welch's *t*-test, $P<10^{-12}$) (Figure 5.7A, red dots; Figure 5.18). Inositol-3-phosphate synthase *INO1* is transcriptionally regulated under stress (235). *INO1* 3' UTR has four conserved crosslinking sites, one of which exhibits >50-fold increase in RBP occupancy upon nitrogen starvation despite a >10-fold decrease in *INO1* mRNA levels (Figure 5.7B; Figure 5.19). These data suggest post-transcriptional regulation of *INO1* mRNA by a specific RBP-RNA interaction in the 3' UTR. We also identified three crosslinking sites under normal growth conditions in the 3' UTR of *AGP3*, an amino acid permease capable of supplying amino acids as an alternative nitrogen source in nitrogen-poor conditions (236). RBP occupancy at these sites was completely lost upon nitrogen starvation while two additional sites emerged (Figure 5.7C; Figure 5.19). *AGP3* mRNA levels

moderately increased ~2- fold (Figure 5.7C), suggesting complex, combinatorial post-transcriptional regulation of *AGP3* expression in nitrogen-poor conditions.

5.4 Discussion

RNP complexes exhibit dynamic properties that are sensitive to environmental conditions. For example, granules containing stalled translation pre-initiation complexes are formed under stress but rapidly dissociate when the cell returns to favorable conditions (237). Despite insight into how particular RNP complexes are affected by stress, global effects of stress on all RBP-RNA interactions have until now remained unexplored. We detect reproducible changes in occupancy for 38% of 3' UTR crosslinking sites on non-translating mRNAs under glucose or nitrogen starvation conditions: loss of RBP occupancy at RBP crosslinking sites was a phenomenon common to both glucose and nitrogen stress conditions, while more distinct sets of crosslinking sites increased RBP occupancy (Figure 5.5C).

In our current work, we limited our gPAR-CLIP analyses to protein-RNA interactions residing in non-translated RNPs (Figure 5.1A, and see Materials and Methods), which mediate important functions for mRNA translation, localization, and degradation. Because we have no information of the identities of the RBPs or their distribution in the sucrose gradient, we cannot distinguish whether the changes in RBP coverage represent changes in RBP binding and/or distribution. This is particularly relevant in glucose or nitrogen starvation, as many RBPs redistribute under these conditions. Future comparative gPAR-CLIP analyses on

both non-translating RNP and translating RNPs in stress conditions will distinguish changes in RBP binding versus changes in RBP localization.

RNAs are capable of forming complex two- and three-dimensional structures, and some RBPs are known to recognize such structural motifs. For example, She2p mediates the localization of several bud-localized transcripts during cell division by recognizing and binding to specific stem-loop structures in mRNAs (226, 227). Examination of the structural properties of our global RBP crosslinking sites revealed a preference for single-stranded regions, which agrees with previous reports of crosslinking sites of the RNA-binding protein FUS occurring at single-stranded regions directly adjacent to the FUS RNA recognition motif (238). Unpaired loop and bulge regions can be unstructured or form tertiary structural modules, both of which can be readily recognized by RBPs. In contrast, double-stranded RNAs, in general, do not provide good platforms for RBP binding: structured RNA regions captured by gPAR-CLIP generally had low CLS values (Figure 5.4) likely resulting from crosslinking and/or RNase T1 cleavage inefficiency. In structured regions, 4-thiouridines are more likely to be locked in U:A or U:G pairing, preventing crosslinking to proteins. In addition, structured regions are less accessible to RNase attack during sequencing fragment preparation, resulting in underrepresentation in gPAR-CLIP libraries. Nevertheless, despite their low crosslinking efficiencies, Ts in double-stranded, paired RNA regions show extremely high conservation compared to Ts with no crosslinking evidence. These data indicate that RNAs with high

secondary structure are evolutionarily conserved and can serve as functional, secondary structure motifs recognized by select RBPs.

RBP binding sites functioning as *cis*-regulatory elements are expected to be under purifying selection. We identified a substantial fraction (35%) of conserved elements in UTRs overlapping RBP crosslinking sites. This represents an underestimation because RBPs and RNAs that are not expressed under our experimental conditions or that fail to crosslink will not be captured. Although crosslinking sites in general are more highly conserved than non-crosslinking sites in UTRs, many sites are not well conserved and might represent species-specific *cis*-regulatory elements that allow adaptation to different environments and stressors.

A preference of RBP binding to 3' UTRs observed in this study and others (67, 68) is consistent with the function and evolution of 3' UTRs as major sites for post-transcriptional regulation. Unlike protein-coding regions, 3' UTRs do not directly engage ribosomes during translation and therefore provide accessible platforms for RBP binding and RNP assembly. One important aspect of gene regulation is combinatorial control, which allows a single gene to be controlled by more than one regulator. In our study, 23% of all nucleotides in annotated 3' UTRs were located within RPB crosslinking sites, corresponding to an average of 1 crosslinking site, on average 23 nt long, in every 100 nt. For a median-sized yeast 3' UTR that is 166 nt long (224), there are on average 2 RBP crosslinking sites, suggesting that most yeast genes are subject to combinatorial post-transcriptional regulation. Since *S. cerevisiae* lacks post-transcriptional regulation

by the highly conserved and pervasive microRNA regulatory pathway, combinatorial regulation by RBPs may play a more prominent role than in organisms with small RNA-mediated post-transcriptional gene regulation.

5.5 Materials and Methods

5.5.1 Strains, media and growth conditions.

The following strains were used in this study: WT BY4742 (*MAT*α *his3*Δ1 *leu2*Δ0 *lys2*Δ0 *ura3*Δ0), TAP-tagged strains picked from TAP-tagged yeast strain collection (*MAT***a** *his3*Δ*1 leu2*Δ*0 met15*Δ*0 ura3*Δ*0 YFG-TAP::HIS5)* (239). Strains were grown at 30°C with vigorous shaking (250 rpm) in synthetic defined media (SD), supplemented with 200 μM 4-thiouracil (4sU, Sigma-Aldrich 440736), to OD_{600} = 0.7-0.8. Starvation was performed by pelleting cells for 5 min at 3,000Xg at room temperature, discarding all media, rinsing once with H_2O, and resuspending cells in an equal volume of SD without glucose or nitrogen (supplemented with 200 μM 4sU). Cells were returned to 30°C with shaking for 2 hr. Strains used are defective in uracil synthesis (*ura3*Δ) and readily take up 4sU from the media. Inside the cell, 4sU is converted by Fur1p (uracil phosphoribosyltransferase) to 4-thiouridine monophosphate that can be incorporated during RNA synthesis.

5.5.2 Estimation of 4sU incorporation rates.

4sU incorporation rates were measured as described (240). Briefly, RNA samples isolated from cells grown in the presence or absence of 4sU were dissolved in 100 µL of 12 mM Tris buffer, pH 7, and their A_{260} absorption was adjusted to the same value. A_{330} was measured for both samples using a Q6 quartz cuvette with 1 mm light path in a Thermo Scientific BioMate 3 UV-Vis spectrophotometer. 4sU incorporation rates per kb RNA were calculated as $500X[(A_{330}(+4sU)) – (A_{330}(-4sU)]/A_{260}$. 4sU was incorporated at roughly four 4sU per kilobase of transcript, with little interference with cell growth and only minor changes in gene expression (Figure 5.9).

5.5.3 gPAR-CLIP procedures.

UV crosslinking. 50 mL of mid-log phase cultures (OD_{600} of 0.7-0.8) were pelleted for 5 min at 3,000Xg at room temperature, resuspended in 2 mL of 1XHBSS (Invitrogen 14025) and transferred to a 60 mm cell culture dish (BD Biosciences 353002), placed on ice, and irradiated with 365nm UV at 150 mJ/cm^2 four times using a UVP CL-1000L UV crosslinker. The cells were then pelleted for 2 min at 5,000Xg at 4°C. After removing 1X HBSS, the cells were frozen in liquid nitrogen.

Extract preparation. Crosslinked cells were resuspended in polysome lysis buffer (20 mM HEPES pH 7.5, 140 mM KCl, 1.5 mM $MgCl_2$, 1% Triton X-100, 1X Complete Mini Protease Inhibitor EDTA-free (Roche 1 836 170), 0.2 U/µL SUPERase·In (Invitrogen AM2696)), mixed with ½ volume of acid-washed

glass beads, and lysed by vortexing four times at 4°C, 1 min each with 1 min incubation on ice in between. Cell debris was removed by centrifugation for 5 min at 1,300Xg at 4°C. The supernatant was cleared by 20,000Xg spin for 10 min at 4°C.

Ribosome depletion using sucrose density gradients. 15-50% (w/v) sucrose density gradients were prepared in Beckman polycarbonate centrifugation tubes (11 X 34 mm) by sequentially layering and freezing 0.24 mL of 50%, 41.25%, 32.5%, 23.75% and 15% sucrose dissolved in polysome gradient buffer (20 mM HEPES pH 7.5, 140 mM KCl, 5 mM $MgCl_2$). Gradients were thawed overnight at 4°C before use. 100 µL of clarified lysate was loaded on top of a gradient, centrifuged for 1 hr at 54,000 rpm at 4°C using a TLS-55 rotor in an Optima MAX-E ultracentrifuge (Beckman Coulter). The top 600 µL of the gradient was recovered and supplemented with 2 µL of SUPERase·In (20 U/µL).

Chemical biotinylation and polyA selection. 60 µL of freshly prepared 10mM EZ-Link NHS-SS-Biotin (Pierce 21441) dissolved in dimethylformamide was added to the recovered lysate and incubated on a Nutator for 2 hr at 4°C. 50 µL of 5 M NaCl was added to increase the total salt concentration to 0.5 M. Biotinylated lysate was mixed with 1 mg of oligo$(dT)_{25}$ magnetic beads (NEB S1419S), then incubated on a Nutator for 30 min at 4°C. The beads were washed four times with ice-cold hybridization buffer (10 mM HEPES pH 7.5, 0.5 M NaCl, 1 mM EDTA) and the RNAs were eluted by incubating beads with 500 µL of elution buffer (10 mM HEPES pH 7.5, 1 mM EDTA) and heating at 65°C for

3 min. The eluted sample was transferred to a new tube and mixed with 55 µL of 10XPBS.

Streptavidin binding and RNase T1 digestion. PolyA-selected samples were mixed with 1 mg of streptavidin M280 Dynabeads (Invitrogen 112-05D) and incubated on a Nutator for 30 min at 4°C. The beads were washed three times with 1XPBS, then incubated with 20 µL of 50 U/µL RNase T1 (Fermentas EN0541, 1:20 dilution in 1XPBS) at 22°C for 15 min on an Eppendorf Thermomixer (15 sec shaking at 1,000 rpm followed by a 2 min rest interval), followed by 5 min incubation on ice. Beads were washed twice with wash buffer (1XPBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40), twice with high-salt wash buffer (5XPBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40) and twice with 1XPNK buffer (50 mM Tris pH 7.4,10 mM $MgCl_2$, 0.5% NP-40).

On-bead CIP treatment. Beads were incubated with 20 µL of CIP mix (50 mM Tris pH 7.9, 100 mM NaCl, 10 mM $MgCl_2$, 0.5 U/µL calf intestinal alkaline phosphatase (CIP) (NEB M0290S)) at 37°C for 15 min, with 15 sec shaking at 1,000 rpm followed by a 2 min rest interval on a Thermomixer. After CIP treatment, beads were washed twice with 1XPNK+EGTA buffer (50 mM Tris pH 7.4, 20 mM EGTA, 0.5% NP-40) and twice with 1XPNK buffer.

On-bead 3' DNA linker ligation. Beads were incubated with 20 µL of ligation mix (50 mM Tris pH 7.4, 10 mM $MgCl_2$, 0.5 mM DTT, 2 µM Pre-adenylated 3' DNA linker, 25% PEG-8000, 10 U/µL T4 RNA ligase 2, truncated K227Q (NEB M0351S)) at 16°C overnight (≥16 hr), with 15 sec shaking at 1,000

rpm followed by a 2 min interval on a Thermomixer. After linker ligation, beads were washed three times with 1XPNK+EGTA buffer.

SDS-PAGE and nitrocellulose transfer. Beads were mixed with 12 µL of 1XPNK+EGTA buffer, 3 µL of freshly made 1M DTT and 15 µL of 4X NuPAGE LDS sample buffer (Invitrogen NP0007), and incubated at 70°C for 10 min. Beads were removed, and the supernatant was loaded onto NuPAGE 4-12% Bis-Tris gel (Invitrogen NP0335BOX) and run at 150 V for 35 min. The gel was transferred to Protran BA 85 nitrocellulose membrane (pore size 0.45 µm, Whatman 10402594) using Novex wet transfer at 30 V for 1 hr. A broad band from 31 kDa up to the top of the gel was excised, cut into small pieces, and transfered into a microfuge tube.

RNA isolation and purification. Excised membranes were incubated with 500 µL of 4 mg/mL Proteinase K prepared in 1X PK buffer (100 mM Tris pH 7.5, 50 mM NaCl, 10 mM EDTA) for 20 min at 37°C on a Thermomixer. 500 µL of 7M urea prepared in 1X PK buffer were added to the tube followed by another 20 min incubation at 37°C. The Proteinase K digestion reaction was mixed with 1 mL of Phenol:Chloroform:Isoamyl Alcohol 25:24:1 (Sigma-Aldrich P2069) by vortexing and spun for 5 min at 20,000Xg. The liquid phase was transferred into a new tube, mixed with 125 µL of 3 M NaOAc, 2.5 mL of 100% ethanol and 1 µL of 15 mg/mL glycoblue (Invitrogen AM9516), and precipitated for 2 hr at -80°C. RNAs were collected by centrifugation for 20 min at 20,000Xg at room temperature followed by two washes with cold 75% ethanol.

RNA 5' end phosphorylation. RNA pellets were air-dried briefly, resuspended in 10 µL of PNK mix (70 mM Tris pH 7.6, 10 mM $MgCl_2$, 5 mM DTT, 1 mM ATP, 1 U/µL T4 polynucleotide kinase (NEB M0201S), 1 U/µL SUPERase·In) and incubated at 37°C for 30 min. The reaction was combined with 90 µL of H_2O and 100 µL of Phenol:Chloroform:Isoamyl Alcohol 25:24:1, mixed well and spun for 5 min at 20,000Xg. The liquid phase was mixed with 12.5 µL of 3 M NaOAc, 250 µL of 100% ethanol, 1 µL of 15 mg/mL glycoblue and precipitated for 2 hr at -80°C. RNAs were collected by centrifugation for 20 min at 20,000Xg at room temperature, followed by two washes with cold 75% ethanol.

5' RNA linker ligation. RNA pellets were resuspended in 10 µL of ligation mix (50 mM Tris pH 7.5, 10 mM $MgCl_2$, 10 mM DTT, 1mM ATP, 0.1 mg/mL BSA, 2 µM 5' RNA linker, 1 U/µL T4 RNA ligase (Fermentas EL0021), 1 U/µL SUPERase•In, 10% DMSO) and incubated at 15°C for 2 hr.

RNA size selection. Ligation reaction was terminated by adding 10 µL of 2X formamide gel loading buffer (Invitrogen AM8546G), heated for 2 min at 70°C and then quickly chilled on ice. Samples were loaded onto a 6% TBE UREA gel (Invitrogen EC6865BOX) and run at 150 V for 45 min. After staining with 1X Sybr Gold Stain (Invitrogen S-11494), a gel piece corresponding to 70-90 nt RNA (80-100 nt ssDNA) was excised, crushed, and soaked in 400 µL of 0.3 M NaOAc overnight at room temperature. After removing gel pieces, the solution was combined with 1 mL of 100% EtOH and 1 µL of 15 mg/mL glycoblue and precipitated for 2 hr at -80°C. RNAs were collected by centrifugation for 20 min at

20,000Xg at room temperature, followed by two washes with cold 75% ethanol. After brief drying, RNAs were resuspended in 15 µL of H_2O.

RT-PCR. 10 µL of the ligated RNA was combined with 2 µL of 5 µM RT primer, heated at 65°C for 5 min, and then quickly chilled on ice, and followed by the addition of 1 µL of 10 mM dNTP, 1 µL of 0.1 M DTT, 4 µL of 5X First strand buffer, 1 µL of SUPERase•In (20 U/µL) and 1 µL of SuperScript III Reverse transcriptase (Invitrogen 18080-093, 200 U/µL). RT reaction was kept at 50°C for 45 min, 55°C for 15 min and 90°C for 5 min. A test PCR was performed with 2.5 µL of RT product in 50 µL PCR mix: 1X AccuPrime PCR buffer I, 0.5 µM P5 long primer, 0.5 µM P7 primer, 0.2 µL AccuPirme Taq High Fidelity (Invitrogen 12346-086, 5 U/µL). PCR was carried out with an initial 3 min denaturation at 98°C, followed by 14-22 cycles of 80 sec denaturation at 98°C, 90 sec annealing and extension at 65°C, and termination with a final 5 min extension at 65°C. 15 µL PCR product was collected after 14, 18, and 22 cycles and analyzed on a 10% TBE gel (Invitrogen EC6275BOX) at 150 V for 1 hr to determine the optimal amplification cycles (the lowest cycle number required to generate 96-116 bp amplicons detected by Sybr Gold staining).

Preparation of sequencing libraries. A 50 µL PCR reaction was carried out with the determined cycle number. Amplicons were purified using DNA clean and concentrator-5 (Zymo D4013), run on 10% TBE gels at 150 V for 1 hr and stained with Sybr Gold. A gel piece corresponding to 96-116 bp DNA was excised, crushed, and soaked overnight in 400 µL 0.3 M NaOAc at room temperature. After removing gel pieces, the solution was combined with 1 mL of

100% EtOH and 1 µL of 15 mg/mL glycoblue and precipitated for 2 hr at -80°C. DNAs were collected by centrifugation for 20 min at 20,000Xg at room temperature, followed by two washes with cold 75% ethanol. After brief drying, amplicons were resuspended in 20 µL of H_2O. 5 µL of purified amplicons were used to seed a second round of PCR in 50 µL: 1X AccuPrime PCR buffer I, 0.5 µM Illumina Primer A, 0.5 µM Illumina Primer B, 0.2 µL AccuPirme Taq High Fidelity for 6-12 cycles. Second PCR amplicons were purified with DNA clean and concentrator-5 (Zymo D4013) and sequenced on an Illumina HiSeq 2000 sequencer.

5.5.4 Puf3p PAR-CLIP procedures.

Puf3p PAR-CLIP was performed similarly to gPAR-CLIP with the following modifications. The *PUF3-TAP::HIS5* strain was cultured and UV-crosslinked as in gPAR-CLIP. Cells were lysed in 1XPBS, 0.5% NP-40, 1X Complete Mini Protease Inhibitor, EDTA-free and cleared by sequential spins at 1,300Xg for 5 min and 20,000Xg for 10 min at 4°C. The clarified lysate was passed through a Costar Spin-X filter (Corning CLS8160), mixed with RNase T1 (Fermentas EN0541) to 1 U/µL, and incubated at 22°C for 15 min followed by 5 min incubation on ice. The lysate was then directly mixed with IgG magnetic beads (prepared by coupling rabbit IgG (Sigma-Aldrich I5006) to Dynabeads M-270 Expoxy (Invitrogen 143-01)) to pull down Puf3p::TAP. RNase T1 digestion, CIP treatment, and 3' DNA linker ligation were performed as described in gPAR-CLIP. Afterwards, 5' end phosphorylation was performed on-bead in 20 µL of

PNK mix (70 mM Tris pH 7.6, 10 mM $MgCl_2$, 5 mM DT, 1 µL P32 rATP (6000 Ci/mmol 10 mCi/mL Perkin Elmer BLU502Z500UC), 1 U/µL T4 polynucleotide kinase) and incubated at 37°C for 15 min. 2 µL of 10 mM ATP was added to the mix and the reaction was incubated for 10 min. After SDS-PAGE and transfer, crosslinked RNAs were visualized by autoradiography and the corresponding Puf3p band was excised. The remaining steps were carried out as described in gPAR-CLIP procedures, omitting the 5' end phosphorylation step.

5.5.5 mRNA-seq procedures.

Yeast strains were grown under normal and starvation conditions described above in the presence of 4sU. Additional replicate mRNA-seq libraries were prepared with yeast strains grown under normal conditions without the additions of 4sU.

Total RNAs were extracted with Acid-Phenol:Chloroform, pH 4.5 with IAA, 25:24:1 (Ambion). Replicate, strand-specific total mRNA-seq libraries, were prepared in parallel using the two linker ligation protocol as described (241).

For preparation of ribo- mRNA-seq libraries, extract preparation, and ribosome depletion using sucrose density gradients were carried out as described in the gPAR-CLIP procedure (avoiding UV crosslinking). PolyA+ mRNAs were enriched using oligo(dT)25 beads and converted into sequencing libraries as described (241).

5.5.6 Oligos for constructing gPAR-CLIP, PAR-CLIP, and mRNA-seq libraries.

Oligos used in this study were synthesized by Integrated DNA Technologies, except the 5' RNA linker, which was synthesized by Dharmacon.

Barcodes: 3' DNA linker oligos (5' phosphorylated, and 3' block with inverted deoxythymindine):

Index 1: 5' pATCACGTCGTATGCCGTCTTCTGCTTGidT 3'

Index 2: 5' pCGATGTTCGTATGCCGTCTTCTGCTTGidT 3'

Index 3: 5' pTTAGGCTCGTATGCCGTCTTCTGCTTGidT 3'

Index 4: 5' pTGACCATCGTATGCCGTCTTCTGCTTGidT 3'

Index 5: 5' pACAGTGTCGTATGCCGTCTTCTGCTTGidT 3'

Index 6: 5' pGCCAATTCGTATGCCGTCTTCTGCTTGidT 3'

Index 7: 5' pCAGATCTCGTATGCCGTCTTCTGCTTGidT 3'

Index 8: 5' pACTTGATCGTATGCCGTCTTCTGCTTGidT 3'

Pre-adenylation of 3' DNA linker oligos was performed with Mth RNA ligase (5' DNA adenylation kit, NEB E2610S) following the vendor's instructions.

5' RNA linker:

5' GUUCAGAGUUCUACAGUCCGACGAUC 3'

Barcoded RT primers:

Index 1: 5' CAAGCAGAAGACGGCATACGACGTGAT 3'

Index 2: 5' CAAGCAGAAGACGGCATACGAACATCG 3'

Index 3: 5’ CAAGCAGAAGACGGCATACGAGCCTAA 3’

Index 4: 5’ CAAGCAGAAGACGGCATACGATGGTCA 3’

Index 5: 5’ CAAGCAGAAGACGGCATACGACACTGT 3’

Index 6: 5’ CAAGCAGAAGACGGCATACGAATTGGC 3’

Index 7: 5’ CAAGCAGAAGACGGCATACGAGATCTG 3’

Index 8: 5’ CAAGCAGAAGACGGCATACGATCAAGT 3’

P7 primer:

5’ CAAGCAGAAGACGGCATACGA 3’

P5 long primer:

5’ AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA 3’

Illumina primer A:

5’ AATGATACGGCGACCACCGA 3’

Illumina primer B:

5’ CAAGCAGAAGACGGCATACGA 3’

5.5.7 Data processing of Illumina HiSeq sequencing reads.

gPAR-CLIP, PAR-CLIP, and mRNA-seq reads were processed to remove linkers, sorted into libraries based on 6-nt barcodes, and removed if they were low quality (Figure 5.8). First, reads with a perfect match to a barcode were

successfully sorted, followed by reads with 1 mismatch to a barcode. If both barcodes were perfectly matched in a read or both barcodes were found with 1 mismatch, the 3'-most barcode was chosen. 99.97% of reads were successfully sorted into libraries under these rules. Next, reads were removed if they met any of the following criteria: <18 nt, only homopolymer As, missing 3' adapter, 5'-3' adapter ligation products, 5'-5' adapter ligation products, low quality (more than 4 bases with quality scores below 10 or more than 6 bases with a quality score below 13). 97.0% of gPAR-CLIP, 80.0% of PAR-CLIP, and 99.7% of mRNA-seq reads passed these filters. High quality reads were mapped to the *S. cerevisiae* genome version S288C with Bowtie (242) using the following parameters: -v 3 (map with up to 3 mismatches), -k 275 (map at up to 275 loci), --best, and --strata. Mapped reads were annotated using custom scripts to known genomic elements in the S288C genome (sacCer3, April 2011) including external UTR annotations (243, 244). Figure 5.8 provides read counts at each processing step.

5.5.8 Assessing data reproducibility.

To determine mRNA-seq and gPAR-CLIP replicate library reproducibility, we calculated replicate correlation using normalized read counts (RPM) of each gene. Pearson correlation coefficients for mRNA-seq libraries ranged from 0.984-0.994, while coefficients for gPAR-CLIP libraries ranged from 0.967-0.971. Due to high reproducibility, subsequent measures of read coverage represent averages of two biological replicate libraries. To determine if the addition of 4sU to growth media substantially alters transcription, biological replicate mRNA-seq

libraries were generated from WT yeast grown under normal conditions without the addition of 4sU. These replicate libraries had a Pearson correlation coefficient of 0.988, indicating high reproducibility, and a Pearson correlation coefficient of 0.982 when compared to WT yeast grown in the presence of 4sU.

5.5.9 Calculation of per-nucleotide crosslinking scores.

To measure RBP crosslinking strength, we calculated a crosslinking score for each genomic T position as the RPM coverage from reads with a T-to-C at that position. Because transcript abundance varies greatly, from zero to tens of thousands of copies, T-to-C coverage of crosslinking sites on highly expressed genes would be preferentially higher than T-to-C coverage of crosslinking sites of lowly expressed genes. To avoid this bias, we normalized T-to-C RPMs to length-normalized transcript abundances (reads per million mapped reads per kilobase of transcript, RPKM) from our mRNA-seq libraries. 2% of Ts with T-to-C RPM coverage in gPAR-CLIP libraries were located on genes that lacked mRNA-seq coverage and were thus removed from further analysis. To adjust for the additional kilobase normalization factor using in RPKM, ratios of gPAR-CLIP RPM:mRNA-seq RPKM were multiplied by a factor of 1,000.

5.5.10 Calculation of RBP crosslinking sites.

Generation of read clusters from gPAR-CLIP libraries. All six gPAR-CLIP libraries were aggregated into one large dataset to generate read clusters. A read cluster was defined as a continuous stretch of nucleotides covered by at

least one gPAR-CLIP read harboring 1 or 2 T-to-C conversion events. This step resulted in 84,136 gPAR-CLIP clusters and 1,915 Puf3p PAR-CLIP clusters.

Defining crosslinking site boundaries. Manual inspection of read clusters revealed long (>100 nt) regions covered by gPAR-CLIP reads containing one or more distinct peaks indicative of distinct crosslinking sites. To distinguish between read peaks within long read clusters and trim low read coverage surrounding strong single peaks, we fit a Gaussian smoothed curve (normal kernel function, bandwidth 21) to each read cluster and used the inflection points of this curve to define the boundaries of individual crosslinking sites. This step resulted in 91,290 gPAR-CLIP crosslinking sites and 1,915 Puf3p PAR-CLIP crosslinking sites.

Calculating read coverage of crosslinking sites. From the set of RBP crosslinking sites derived from all gPAR-CLIP libraries, we determined read coverage for each site from each individual library by calculating the average RPM covering each nucleotide in the crosslinking site. This coverage was divided by the RPKM of the associated gene and multiplied by 1000 to enable direct comparison of RBP occupancy of crosslinking sites between growth conditions.

Assigning FDR to each crosslinking site. A small fraction of T-to-C mismatches in gPAR-CLIP reads likely represent sequencing error instead of crosslinking events, so crosslinking sites derived from this error were removed. We repeated the crosslinking site generation steps using mRNA-seq reads with 1 or 2 T-to-C mismatches, which represent the rate of T-to-C sequencing error for the Illumina HiSeq platform. For each gPAR-CLIP and mRNA-seq crosslinking

site, we calculated the T-to-C conversion rate as the number of reads with T-to-C conversion events divided by the number of total reads covering Ts. gPAR-CLIP and mRNA-seq crosslinking sites were binned into groups based on total read coverage. For each gPAR-CLIP crosslinking site in each bin, we determined the proportion of mRNA-seq crosslinking sites with a higher T-to-C coversion rate than the gPAR-CLIP crosslinking site. This proportion represents the false discovery rate (FDR) for that gPAR-CLIP crosslinking site. Using a strict 1% FDR threshold, we identify 80,883 gPAR-CLIP crosslinking sites.

5.5.11 Effect of counting statistics on error in crosslinking site coverage measurement.

Read coverage of replicate gPAR-CLIP crosslinking sites were analyzed to measure reproducibility. For each site, we compared the number of reads coming from one replicate library to the total number of reads from both libraries. Perfect reproducibility would result in a ratio of ½. We binned crosslinking sites based on total RPM and calculated the standard deviation of these ratios for each bin. We predicted the standard deviation for counting statistics by binomial partitioning of total reads for each crosslinking site in each bin between the two replicates. When the total number of reads was below 5 RPM, binomial partitioning predominantly contributed to replicate variation (Figure 5.14). Above 5 RPM, replicate variation stabilized, and counting statistics error contributed little to replicate error.

5.5.12 Conservation analysis.

phastCons conservation scores for each genomic nucleotide were downloaded from Siepel *et al*. (225). Ts with CLSs were grouped into 5' UTR, CDS, and 3' UTR regions and then ranked and binned by CLS so each bin overlapped adjacent bins by 50%. phastCons scores in each bin were averaged. As controls, Ts with no CLS were grouped in 5' UTR, CDS, and 3' UTR regions, randomly ranked, and binned as described. phastCons scores in each bin were averaged. Controls were calculated ten times for each region.

5.5.13 Unpaired probability analysis.

The unpaired probability of each genomic position was calculated using RNAplfold (228) from the ViennaRNA package version 1.8.5 using a span of 40 nt and an averaging window of 80 nt. Ts with CLSs were grouped into 5' UTR, CDS, and 3' UTR regions and then ranked and binned by CLS so each bin overlapped adjacent bins by 50%. Unpaired probabilities in each bin were averaged. As controls, Ts with no CLS were grouped into 5' UTR, CDS, and 3' UTR regions, randomly ranked, and binned. The unpaired probabilities in each bin were averaged.

5.5.14 Crosslinking site pairedness analysis.

Genomic regions corresponding to crosslinking sites were extended to 80 nt centered on the original crosslinking site. These sequences were subjected to folding using RNAfold (245) from the ViennaRNA package version 1.8.5, and the

minimum free energy structures were extracted. Predicted structures were aligned and ranked by average crosslinking site CLS and divided into 100 equally sized, non-overlapping bins. The percentage of nucleotides predicted to be unpaired at each position in each bin was computed. Selected structures from low, middle, and high CLS bins were visualized using VARNA (246).

5.5.15 Enriched motif analysis.

gPAR-CLIP crosslinking sites passing a 5% FDR threshold from genes identified as RBP targets by RIP-Chip experiments (66, 67) were analyzed by MEME (247). 5' UTRs, CDS, and 3' UTRs crosslinking sites were analyzed separately, and third-order Markov models based on all 5' UTR, CDS, or 3' UTR regions were used to model background nucleotide compositions. Because gPAR-CLIP crosslinking sites on each target might represent a combination of RBP recognition sites, we implemented MEME using the –mods zoops parameter to allow zero or one motif to be found in each site. The following parameters were also used: -evt 20, -minw 6, and -maxw 15.

5.5.16 Gene Ontology enrichment analysis.

GO analysis was performed on genes harboring 3' UTR crosslinking sites that were 4-fold up- or down-regulated upon glucose or nitrogen starvation or both. The topGO R Bioconductor package was implemented using Fisher's exact test for enrichment and Bonferroni correction of p-values to adjust for multiple testing (248). Up to 20 GO terms were reported with a p-value <0.01.

5.6 Acknowledgments

We thank L. Weisman and D. Klionsky for yeast strains and discussions. We also thank A. Billi, N. Weiser, and K. Gunsalus for comments on the manuscript.

Table 5.1 Sequencing and mapping statistics.

condition	protocol	4sU added?	Ribosome depletion?	total raw reads	total high quality reads	total mapped reads			total mapped reads (1 T>C)	% of high quality reads	total mapped reads (2 T>C)	% of high quality reads
						0 mm	1 mm	2 mm				
WT	gPAR-CLIP	Yes	Yes	10,345,660	9,892,996	855,723	5,652,529	1,080,794	5,396,767	54.6%	692,138	7.0%
WT	gPAR-CLIP	Yes	Yes	13,354,315	12,354,153	1,020,368	7,138,708	1,184,089	6,848,424	55.4%	728,086	5.9%
-glucose	gPAR-CLIP	Yes	Yes	8,698,657	8,571,862	1,186,644	4,130,861	683,064	3,808,408	44.4%	275,737	3.2%
-glucose	gPAR-CLIP	Yes	Yes	11,545,105	11,415,553	1,587,380	5,253,564	973,283	4,828,045	42.3%	400,089	3.5%
-nitrogen	gPAR-CLIP	Yes	Yes	12,565,700	12,315,329	1,846,036	6,128,110	837,766	5,654,249	45.9%	274,345	2.2%
-nitrogen	gPAR-CLIP	Yes	Yes	7,524,557	7,393,201	1,009,996	3,616,734	508,681	3,334,758	45.1%	174,574	2.4%
WT	mRNA-seq	Yes	No	11,164,331	11,126,176	8,074,185	823,549	92,132	341,800	3.1%	9,700	0.1%
WT	mRNA-seq	Yes	No	11,626,588	11,596,323	8,495,022	888,394	105,960	345,756	3.0%	9,539	0.1%
WT	mRNA-seq	Yes	Yes	20,881,199	20,881,199	15,528,464	161,166	119,230	47,856	0.2%	2,241	0.0%
WT	mRNA-seq	Yes	Yes	19,254,100	19,254,100	14,955,841	145,716	110,676	47,400	0.2%	3,927	0.0%
WT	mRNA-seq	No	Yes	18,570,981	18,570,981	14,911,218	206,401	157,157	34,727	0.2%	2,567	0.0%
WT	mRNA-seq	No	Yes	18,618,595	18,618,595	14,654,128	140,281	112,501	22,707	0.1%	2,185	0.0%
-glucose	mRNA-seq	Yes	Yes	20,764,203	20,764,203	17,367,447	143,677	114,825	26,350	0.1%	2,251	0.0%
-glucose	mRNA-seq	Yes	Yes	16,396,102	16,396,102	13,415,811	115,996	92,410	22,998	0.1%	2,518	0.0%
-nitrogen	mRNA-seq	Yes	Yes	17,424,942	17,424,942	14,094,672	145,566	102,798	31,929	0.2%	2,246	0.0%
-nitrogen	mRNA-seq	Yes	Yes	24,514,654	24,514,654	19,641,672	220,397	157,438	42,438	0.2%	3,057	0.0%
WT (Puf3p IP)	PAR-CLIP	Yes	No	10,498,429	8,394,537	1,450,253	3,431,211	470,044	3,219,759	38.4%	181,296	2.2%

Read counts and T-to-C conversion rates for all gPAR-CLIP, mRNA-seq, and PAR-CLIP libraries.

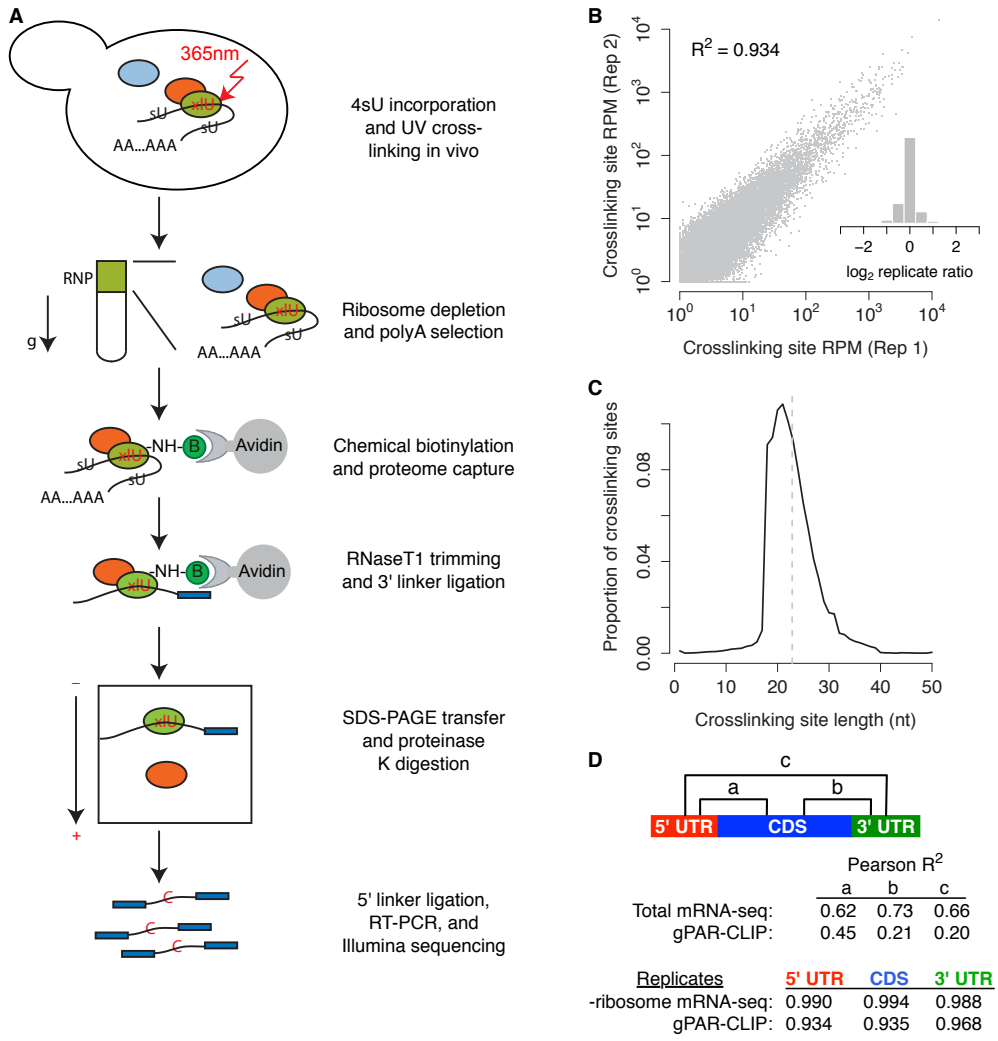


Figure 5.1 gPAR-CLIP identifies transcriptome-wide RBP crosslinking sites. (A) Schematic of the gPAR-CLIP protocol. (B) Reproducibility of crosslinking sites generated from replicate gPAR-CLIP libraries prepared from yeast grown in synthetic defined media (abbreviated as WT gPAR-CLIP hereafter). Pearson correlation coefficient is indicated. Inset: distribution of $\log_2$ crosslinking site RPM ratios between replicates. Replicate error σ=1.3-fold. (C) Length distribution of crosslinking sites in WT gPAR-CLIP libraries. Dotted line: average crosslinking site length of 23 nt. (D) Pearson correlation coefficients of total mRNA-seq and gPAR-CLIP read coverage between 5' UTR, CDS, and 3' UTR regions as well as correlation coefficients of ribosome depleted (-ribosome) mRNA-seq and gPAR-CLIP read coverage between replicate WT libraries.

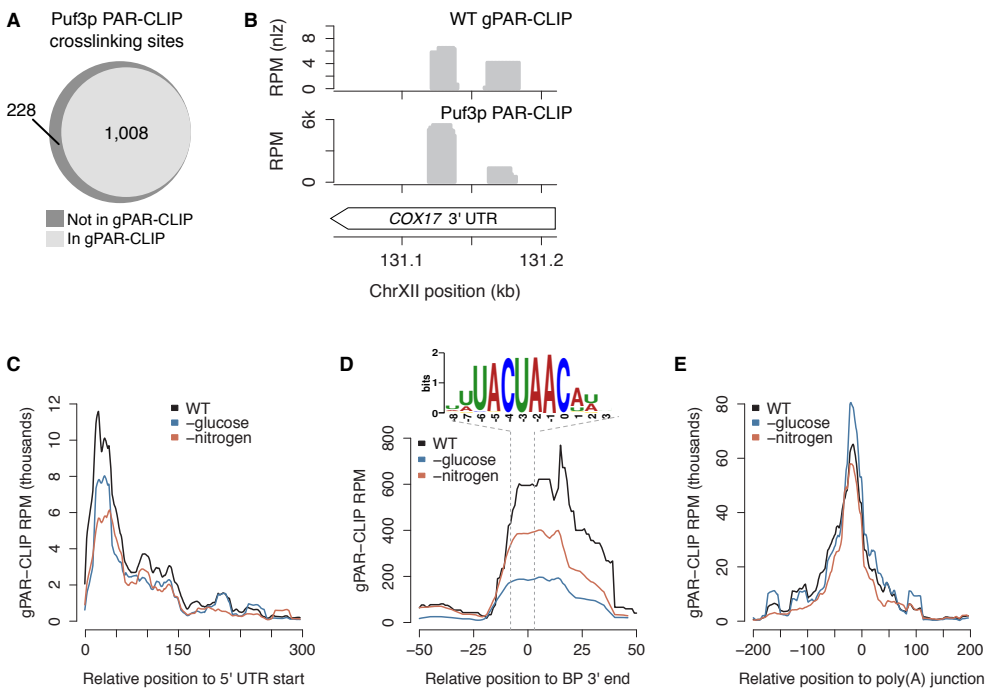


Figure 5.2 gPAR-CLIP captures known RBP crosslinking signatures.
(A) Overlap of crosslinking sites identified in Puf3p PAR-CLIP and WT gPAR-CLIP. Puf3p PAR-CLIP crosslinking sites with >1% T-to-C conversion rate (see Figure 5.11) were considered captured by gPAR-CLIP if at least 50% of their nts overlapped with a WT gPAR-CLIP crosslinking site with FDR<1%. (B) Identification of known Puf3p binding sites on *COX17* mRNA in WT gPAR-CLIP and Puf3p PAR-CLIP. (C-E) Aggregate gPAR-CLIP crosslinking site coverage of the first 300 nt of 2,626 annotated 5' UTRs (C), 51 annotated ribosomal gene introns centered at the branch point (BP) 3' end (D), and 4,241 3' UTRs centered on the poly(A) junction (E).

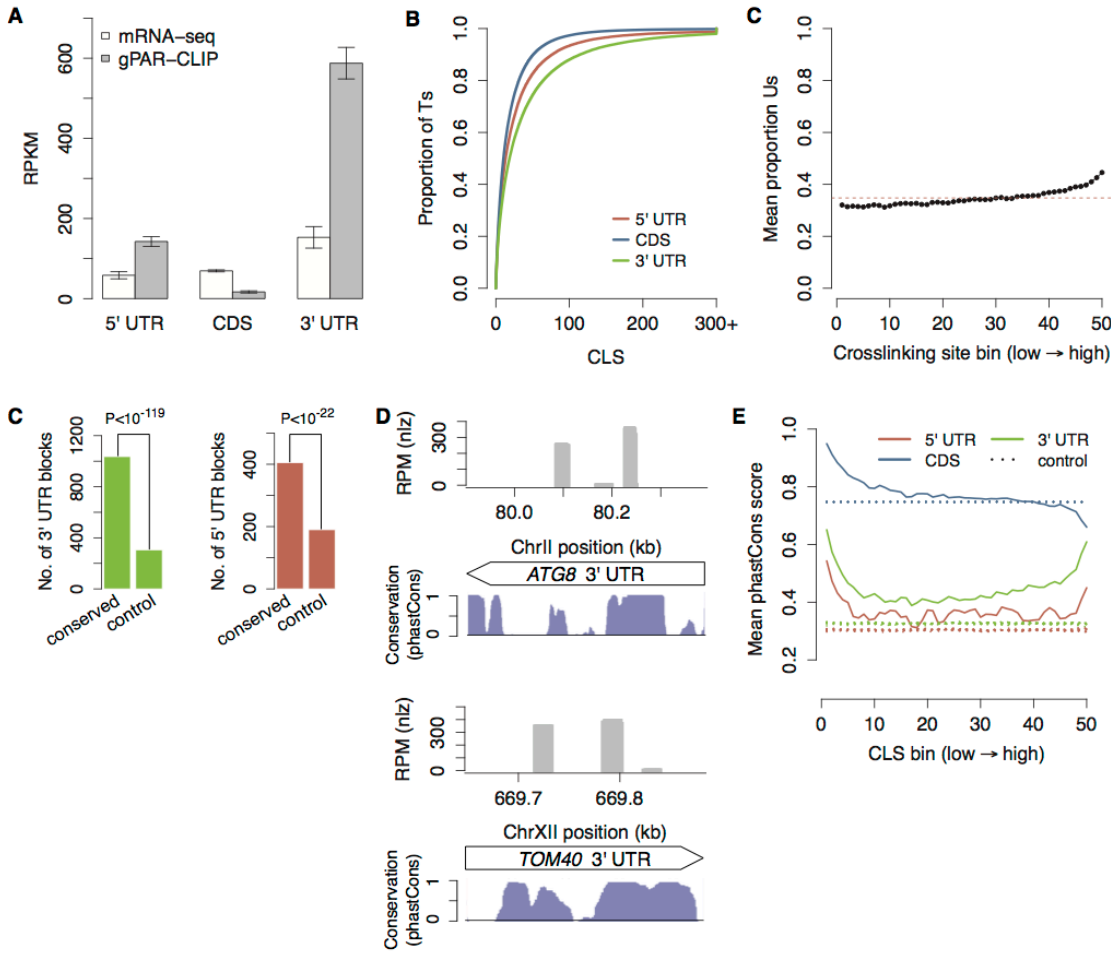


Figure 5.3 RBP crosslinking sites exhibit global sequence conservation.
(A) Average ribosome depleted mRNA-seq and gPAR-CLIP read distributions across 5' UTR, CDS, and 3' UTR regions for all libraries. Error bar: 1 standard deviation. RPKM: reads per million mapped reads per kilobase. (B) Cumulative distribution of CLS values from WT libraries. (C) Proportion of Ts in crosslinking site binned by crosslinking site coverage (RPM). Dotted red line indicates average T content of all crosslinking sites. (D) Number of conserved blocks in 3' and 5' UTRs overlapping 100% with WT gPAR-CLIP crosslinking sites (χ^2 p-values indicated). Control blocks were randomly generated within 3' and 5' UTRs to match the number and size of conserved blocks. (E) Two major gPAR-CLIP crosslinking sites in *ATG8* 3' UTR (top) and *TOM40* 3' UTR (bottom) overlapping conserved blocks. (F) Mean phastCons scores for Ts ranked and binned by CLS. Control lines represent mean phastCons scores of randomly ranked and binned Ts with no CLS, repeated 10 times.

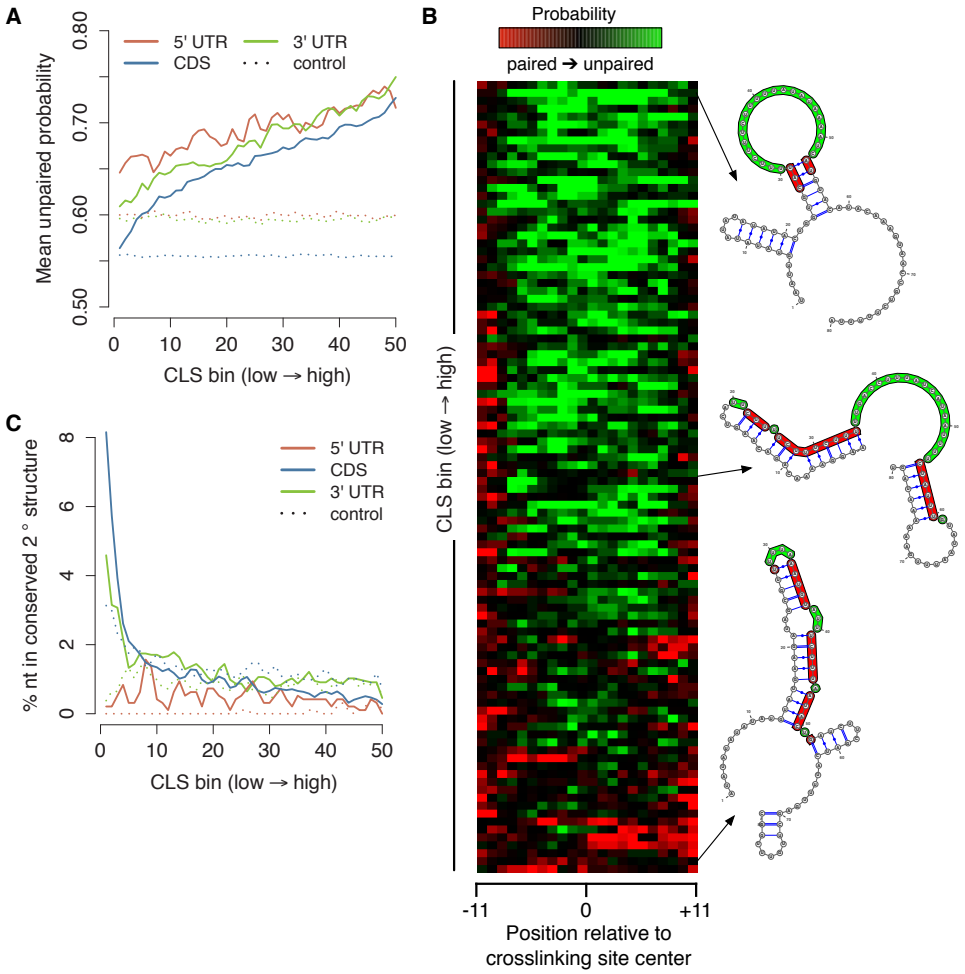


Figure 5.4 RBP crosslinking sites share global structural characteristics. (A) Mean unpaired probability scores for Ts ranked and binned by CLSs. Control lines represent mean unpaired probability of randomly ranked and binned Ts with no CLS. Pearson correlation coefficients: 5' UTR R^2=0.933, CDS R^2=0.976, 3' UTR R^2=0.986. (B) Crosslinking site pairedness visualized as a heatmap. Columns represent nucleotide positions within crosslinking sites. Rows represent average unpaired probability for 100 crosslinking sites in that bin. Select secondary structure predictions from low, middle, and high CLS regions are indicated with the crosslinking site colored. (C) Percentage of Ts ranked and binned by CLSs in conserved secondary structural elements as defined by RNAz. Control lines represent percentage of randomly ranked and binned Ts with no CLS in conserved secondary structural elements.

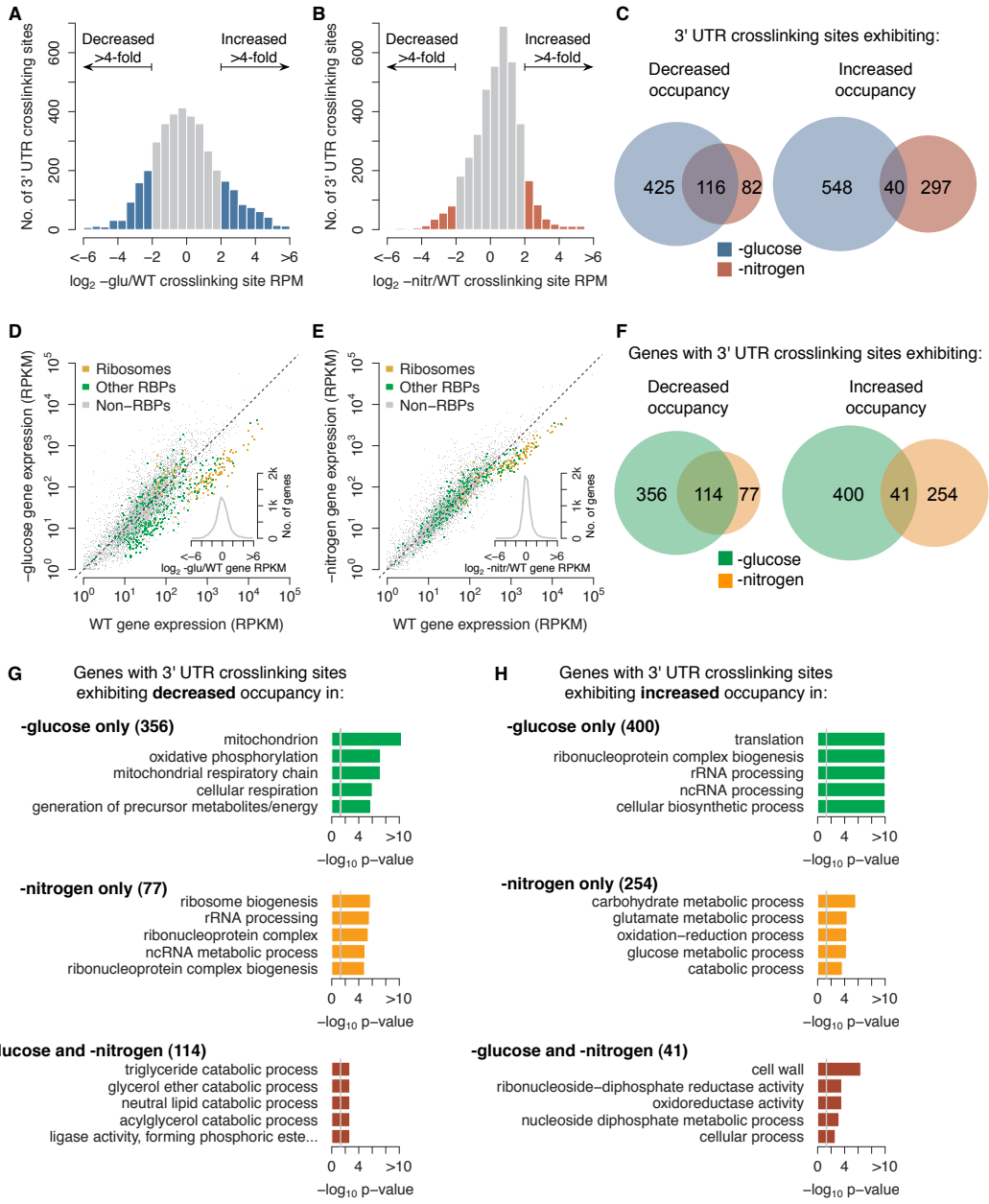


Figure 5.5 Nutrient deprivation induces global but distinct RBP-crosslinking and mRNA changes.

(A-B) Global changes in 3' UTR crosslinking site coverage upon glucose (A) or nitrogen (B) starvation. Standard deviations of intra-replicate variation: WT 1.31-fold; glucose starvation 1.24-fold; nitrogen starvation 1.15-fold. (C) Overlap of 3' UTR crosslinking site changes affected by glucose or nitrogen starvation conditions. (D-E) Global changes in mRNA abundance upon glucose (D) or nitrogen (E) starvation. (F) Overlap of mRNAs with 3' UTR crosslinking site changes affected by glucose or nitrogen starvation conditions. (G-H) Enriched GO terms for mRNAs with 3' UTR crosslinking sites with decreased (G) or

increased (H) RBP occupancy upon glucose or nitrogen starvation or both. Grey lines indicate p-value of 0.05.

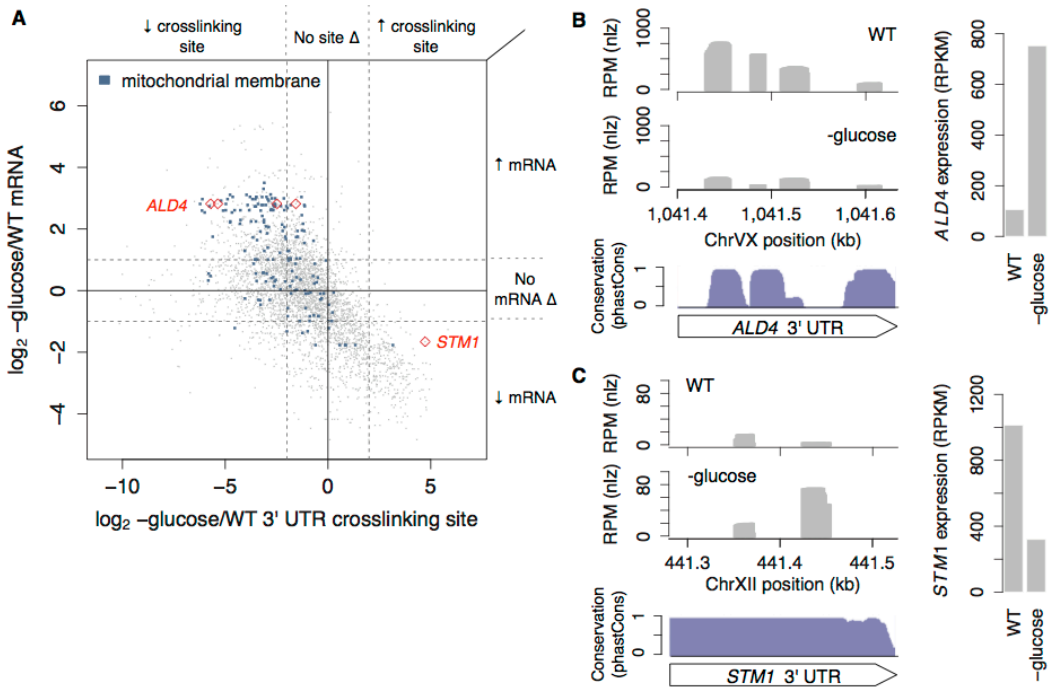


Figure 5.6 Glucose starvation induces RBP-crosslinking and mRNA changes associated with mitochondrial processes.
(A) Global changes in 3' UTR crosslinking site coverage versus changes in the corresponding mRNA upon glucose starvation. Crosslinking sites on genes annotated with "mitochondrial membrane" GO term are colored blue. Dotted lines indicate ≥4-fold changes in crosslinking site coverage (vertical) or ≥2-fold change in mRNA expression (horizontal). (B) *ALD4* 3' UTR contains four crosslinking sites that decrease 2- to 8-fold in RBP occupancy upon glucose starvation and overlap with conserved blocks (red diamonds in (A)). *ALD4* mRNA expression is up-regulated upon glucose starvation. (C) *STM1* 3' UTR contains one crosslinking site that increases in coverage upon glucose starvation (red diamond in (A)). *STM1* mRNA expression is down-regulated upon glucose starvation.

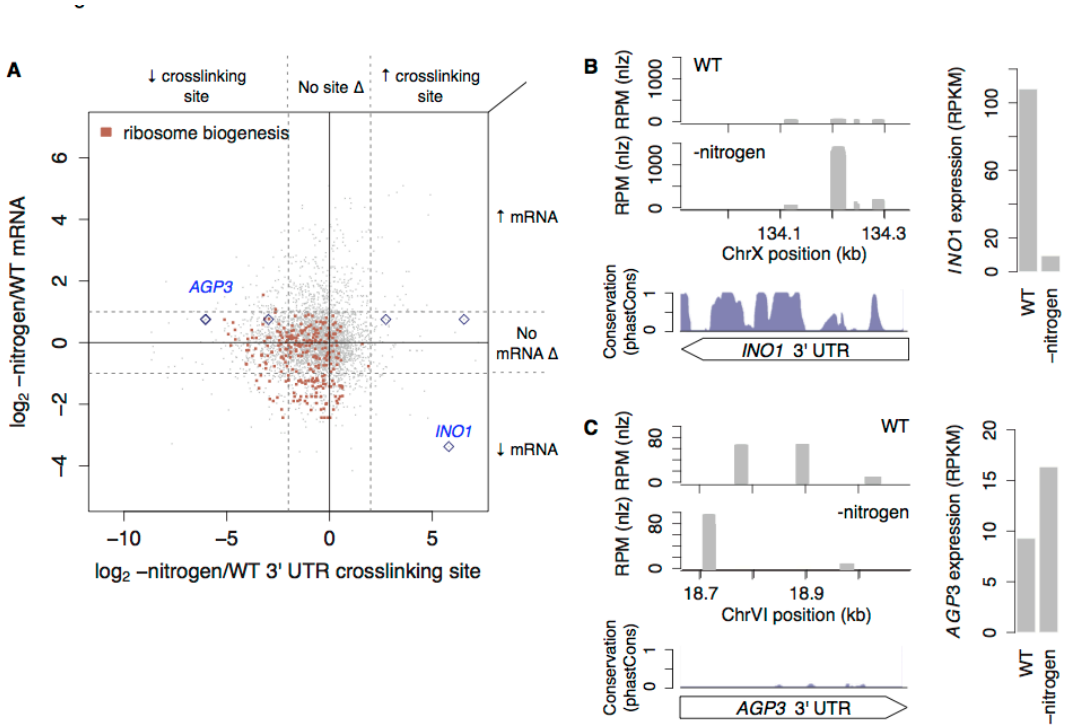


Figure 5.7 Nitrogen starvation induces specific RBP-crosslinking and mRNA changes associated with ribosomes and translation-related processes.

(A) Global changes in 3' UTR crosslinking site coverage versus changes in the corresponding mRNA upon nitrogen starvation. Crosslinking sites on genes annotated with "ribosome biogenesis" GO term are colored red. Dotted lines indicate ≥4-fold changes in crosslinking site coverage (vertical) or ≥2-fold change in mRNA expression (horizontal). (B) *INO1* 3' UTR contains one crosslinking site that increases in coverage upon nitrogen starvation and falls within a conserved block (blue diamond in (A)). *INO1* mRNA expression is down-regulated upon nitrogen starvation. (C) *AGP3* 3' UTR contains three crosslinking sites that are lost and two crosslinking sites that appear upon nitrogen starvation (blue diamonds in (A)). *AGP3* mRNA expression is up-regulated upon nitrogen starvation.

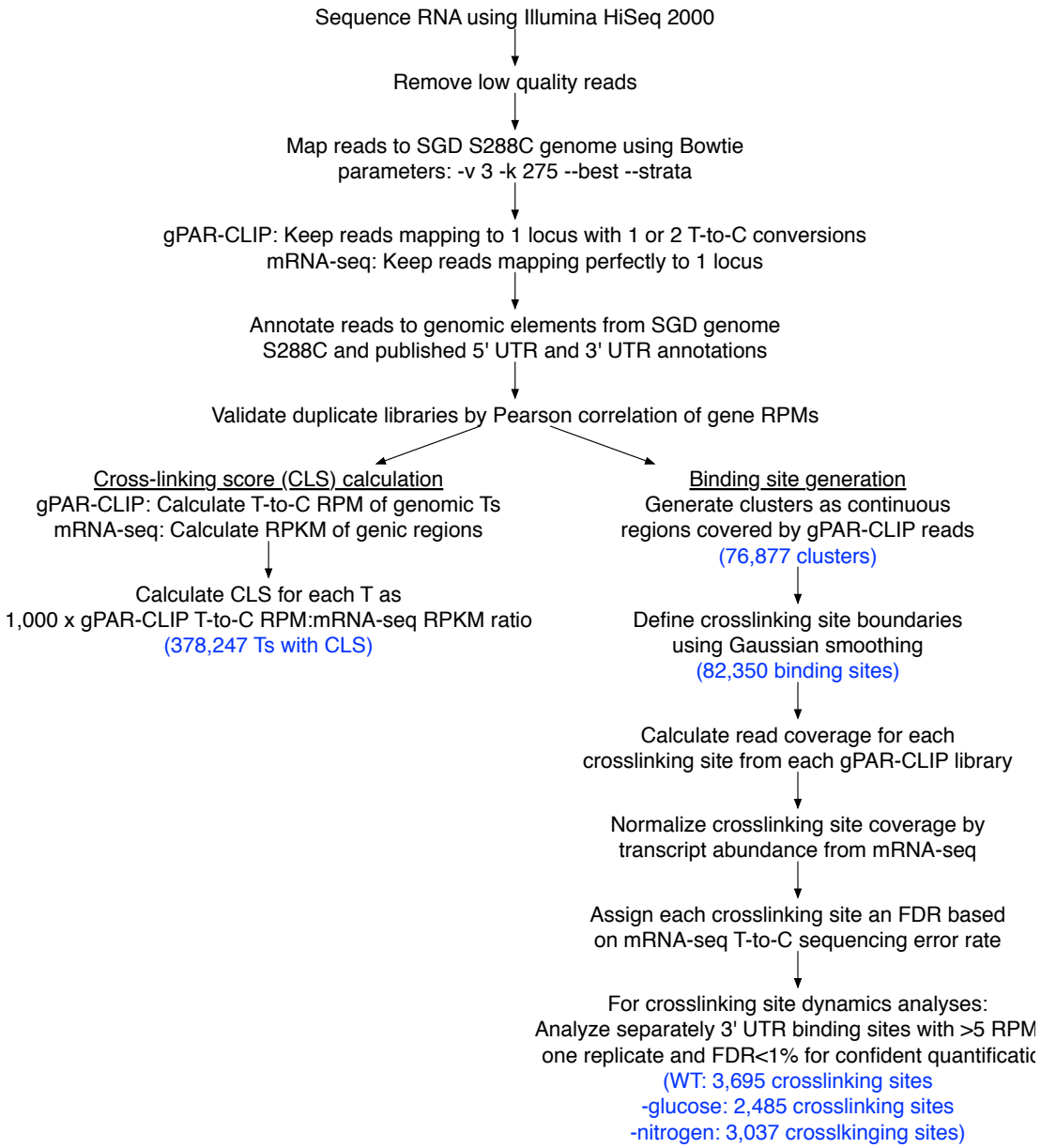
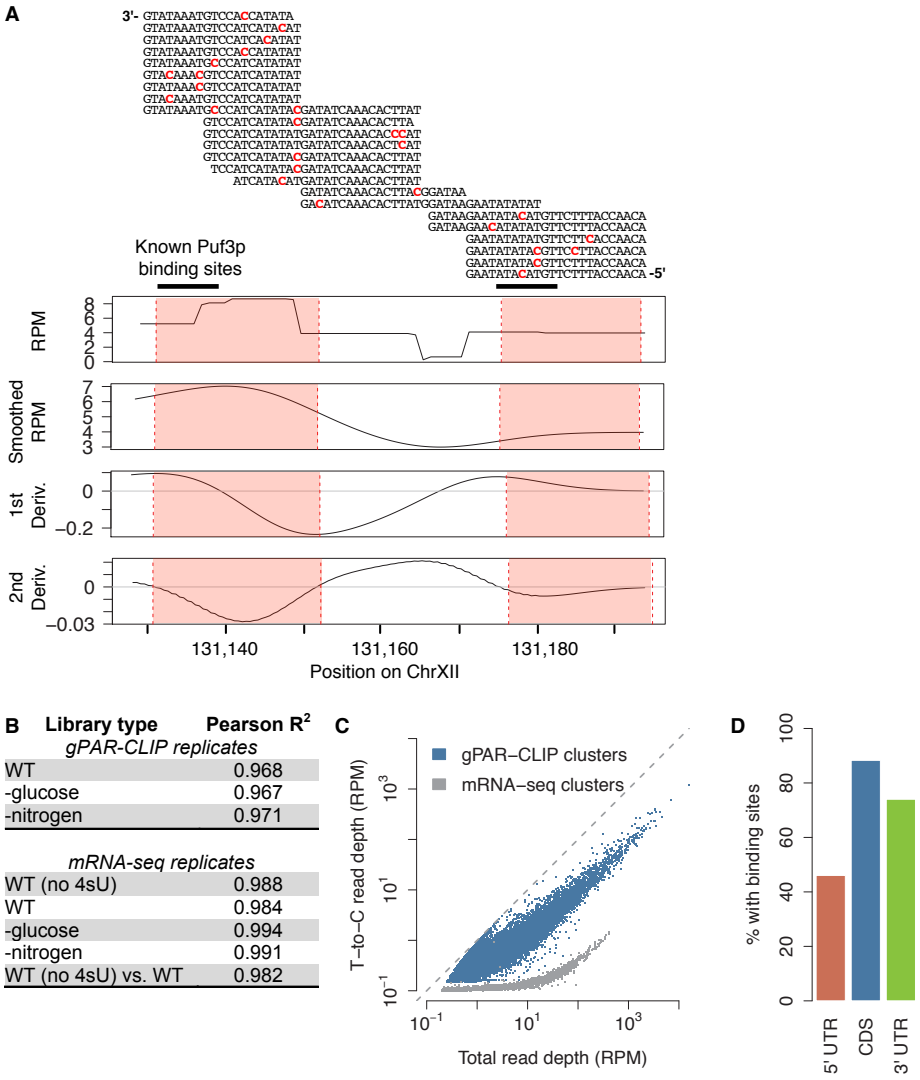


Figure 5.8 Pipeline for generating crosslinking scores and crosslinking sites.
Processing steps used to generate crosslinking scores and crosslinking sites from gPAR-CLIP and mRNA-seq data.



Library type	Pearson R^2
gPAR-CLIP replicates	
WT	0.968
-glucose	0.967
-nitrogen	0.971
mRNA-seq replicates	
WT (no 4sU)	0.988
WT	0.984
-glucose	0.994
-nitrogen	0.991
WT (no 4sU) vs. WT	0.982

Figure 5.9 Computational identification of crosslinking sites.
(A) Illustration of sequence block generation, Gaussian distribution fitting, and cluster segmentation to identify individual crosslinking sites. (B) Pearson correlation coefficients for all gPAR-CLIP and mRNA-seq replicate libraries based on gene RPM values. (C) Separation of T-to-C sequencing errors from crosslinking-induced mismatches. Plotted for each cluster is T-to-C RPM coverage versus total RPM coverage from gPAR-CLIP or mRNA-seq libraries. (D) Percent of annotated 5' UTR, CDS, and 3' UTR regions with at least 1 crosslinking site with >5 RPM.

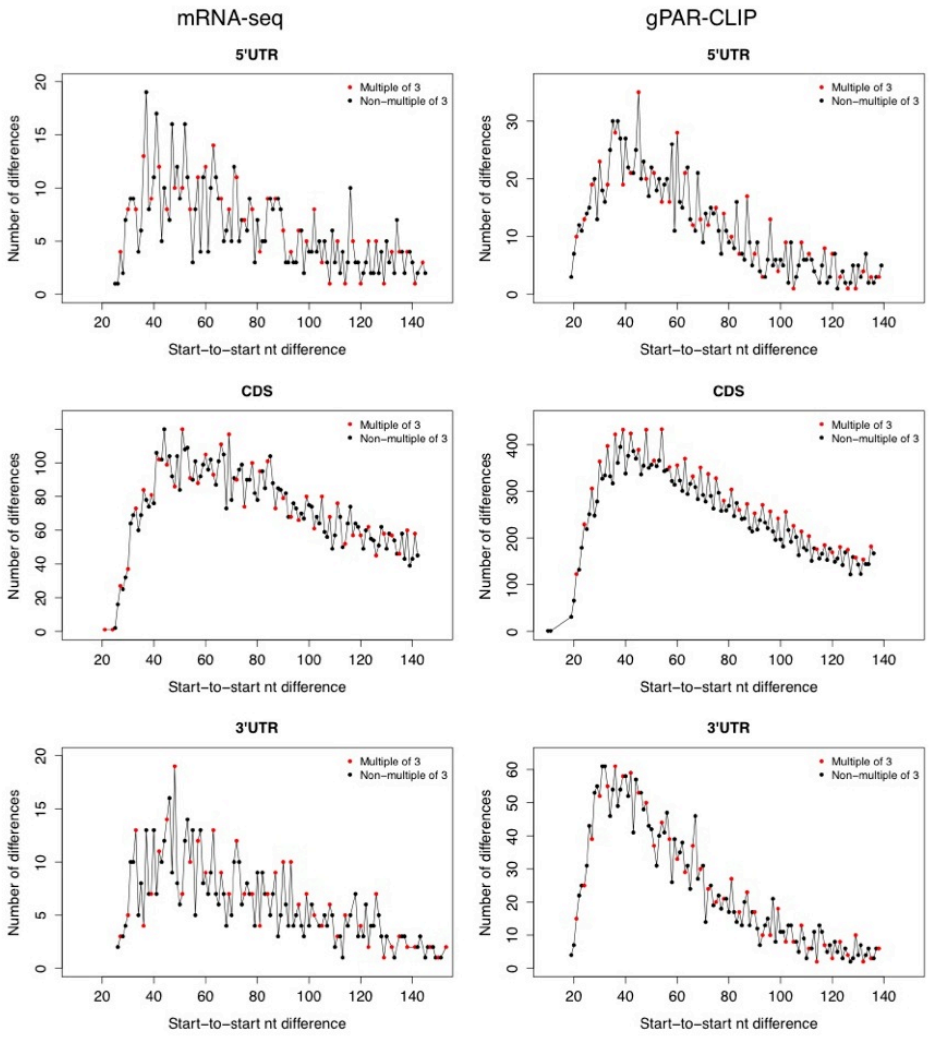


Figure 5.10 Visualization of crosslinking site periodicity.
Distribution of start-to-start nucleotide distances between 5' UTR, CDS, and 3' UTR read clusters from gPAR-CLIP and mRNA-seq libraries. Only distances from gPAR-CLIP CDS read clusters were enriched for multiples of 3 (red dots).

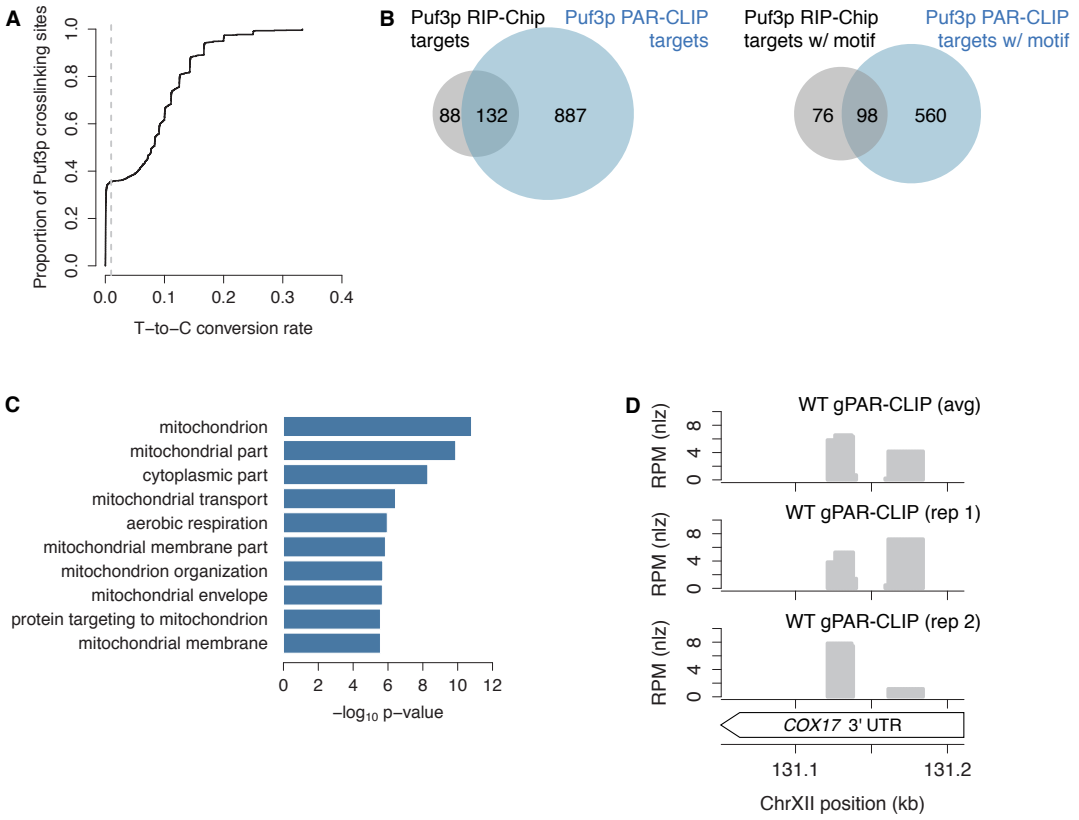


Figure 5.11 Analysis and comparison of PAR-CLIP-identified Puf3p targets. (A) Puf3p PAR-CLIP identified crosslinking sites in 147 (67%) of the 220 Puf3p target mRNA identified by RIP-Chip. 174 Puf3p RIP-Chip-identified target mRNAs contain the Puf3p recognition motif UGUAAAUA. Puf3p PAR-CLIP identified motif-containing crosslinking sites in 76 (44%) of these mRNAs and in 265 additional mRNAs, suggesting post-transcriptional regulation by Puf3p for these 265 novel targets. (B) GO enrichment analysis of 265 PAR-CLIP-identified, motif-containing Puf3p targets. Results are consistent with Puf3p's role in localization, deadenylation, and repression of mRNAs encoding proteins destined for the mitochondria. (C) Individual replicate coverage of *COX17* 3' UTR in gPAR-CLIP with average coverage as shown in Figure 5.2B.

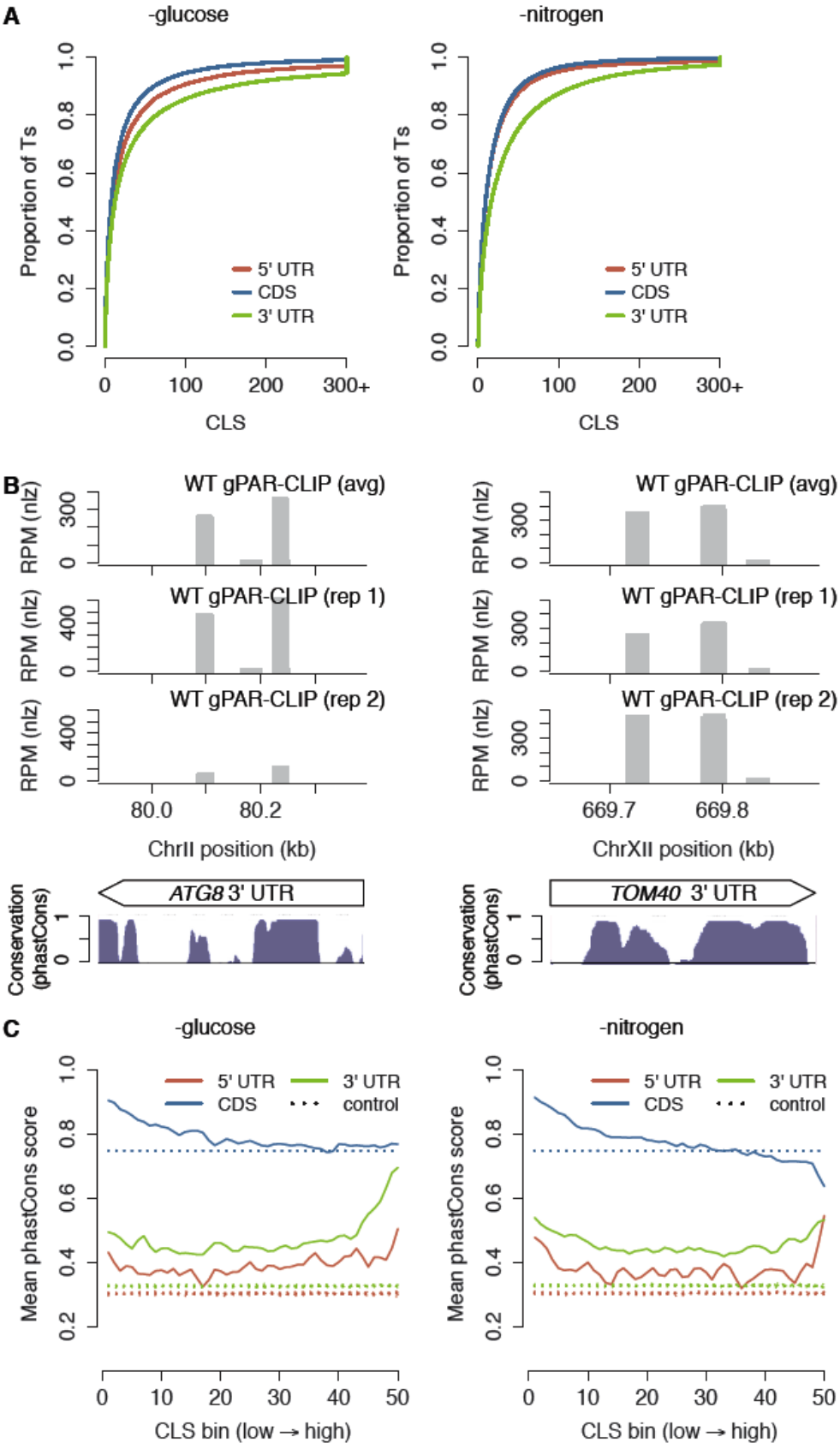

A
-glucose
-nitrogen
Proportion of Ts
0.0
0.2
0.4
0.6
0.8
1.0
5' UTR
CDS
3' UTR
0
100
200
300+
CLS
B
WT gPAR-CLIP (avg)
WT gPAR-CLIP (rep 1)
WT gPAR-CLIP (rep 2)
RPM (nlz)
80.0
80.2
669.7
669.8
ChrII position (kb)
ChrXII position (kb)
ATG8 3' UTR
TOM40 3' UTR
Conservation (phastCons)
C
-glucose
-nitrogen
Mean phastCons score
control
0
10
20
30
40
50
CLS bin (low → high)

Figure 5.12 Analysis of crosslinking scores and conservation of genomic Ts in starvation conditions.
(A) Cumulative distribution of CLSs from 5' UTR, CDS, and 3' UTR regions. (B) Individual replicate coverage of ATG8 and TOM40 3' UTRs in gPAR-CLIP with average coverage as shown in Figure 5.3D. (C) Mean phastCons scores for Ts ranked and binned by CLSs. Control lines represent mean phastCons scores of randomly ranked and binned Ts with no CLS, repeated 10 times.

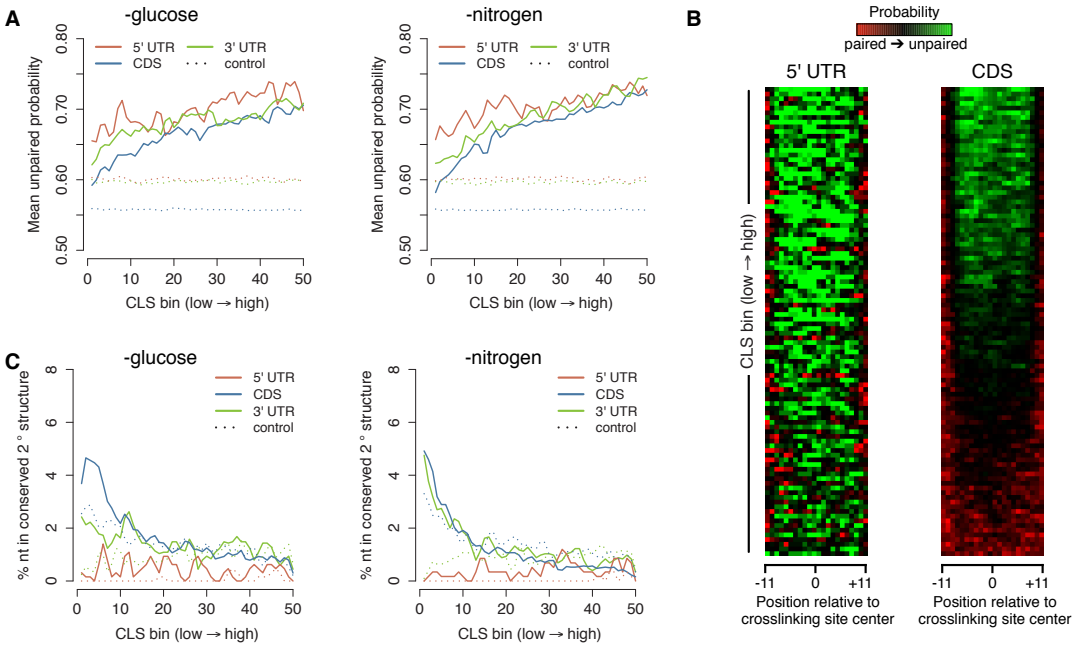


Figure 5.13 Analysis of RNA secondary structure in starvation conditions. (A) Mean unpaired probability scores for Ts ranked and binned by CLSs. Control lines represent mean unpaired probability of randomly ranked and binned Ts with no CLS, repeated 10 times. (B) Heatmaps of pairedness of 5' UTR and CDS crosslinking sites ranked by average crosslinking site CLS. (C) Percentage of Ts ranked and binned by CLSs in conserved secondary structural elements as defined by RNAz. Control lines represent percentage of randomly ranked and binned Ts with no CLS in conserved secondary structural elements, repeated 10 times.

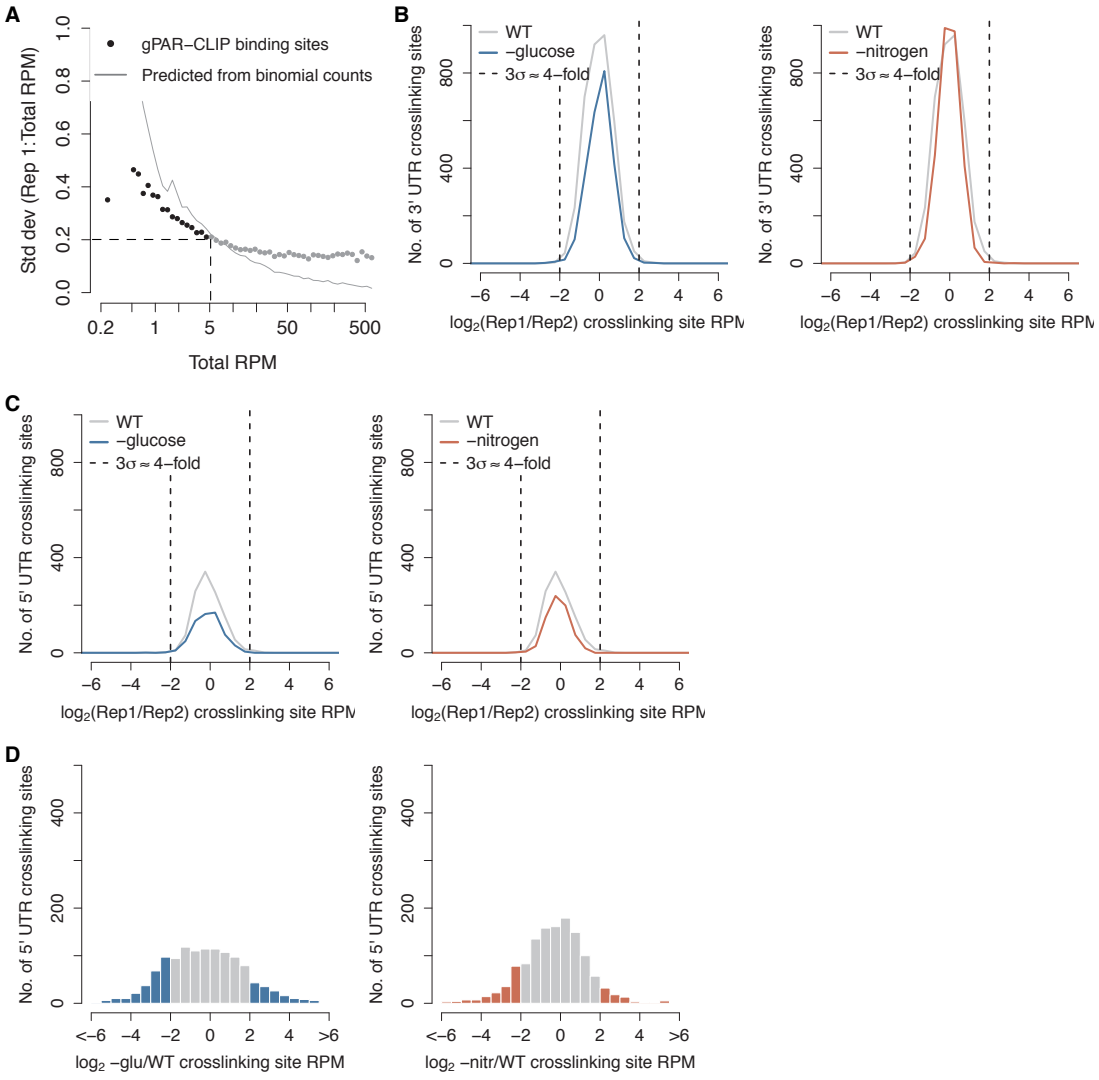
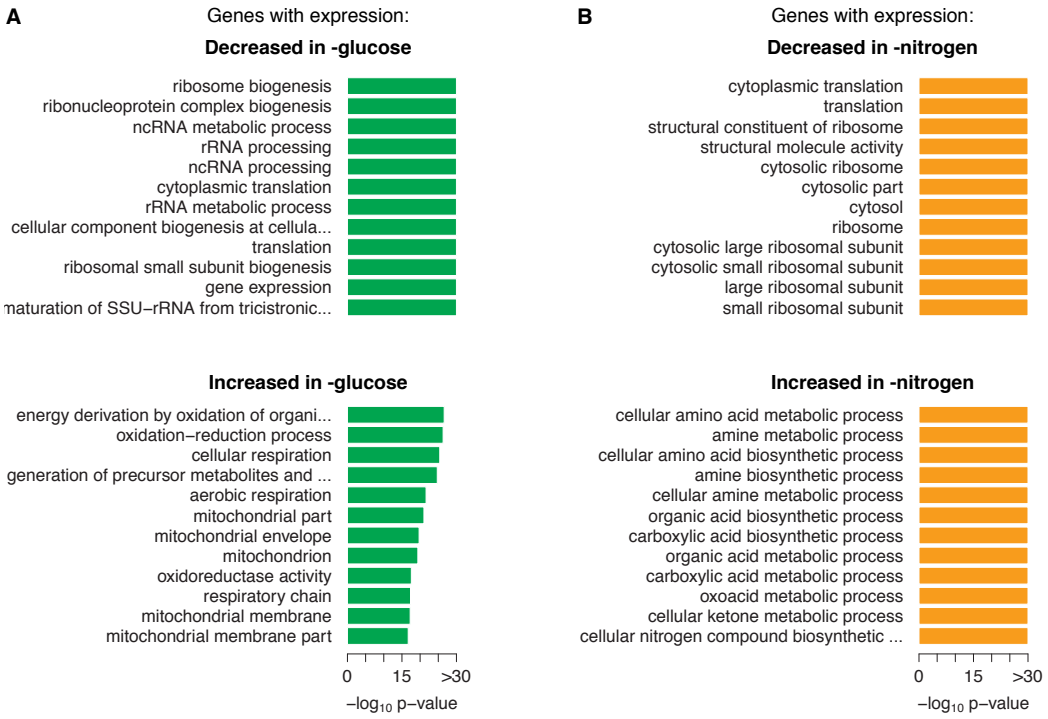


Figure 5.14 Intra-replicate variation of crosslinking site coverage and global changes in 5' UTR crosslinking sites.
(A) Determination of minimum crosslinking site coverage required for comparison of sites across environmental conditions. 5 RPM was chosen as the minimum crosslinking site coverage needed for confident quantification since, at this coverage, the standard deviation of the fraction of crosslinking site reads coming from one replicate library stabilized at <0.2. Shown is data from WT replicate libraries; similar results were obtained for all library types. (B) Intra-replicate variation of 3' UTR crosslinking sites in WT and glucose (left) or nitrogen (right) starvation conditions. Dotted lines represent 3 standard deviations from the mean and correspond to ~4-fold change between WT replicates. (C) Same as (A) but for 5' UTR crosslinking sites. (D) Global changes in 5' UTR crosslinking site coverage upon glucose (left) or nitrogen (right) starvation.



C

-glucose 3' UTR crosslinking sites (%)

mRNA	*Decreased*	*No change*	*Increased*	*Total*
Increased	363 (10)	1,325 (35)	17(0)	1,705 (45)
No change	174 (5)	1,234 (32)	330 (9)	1,738 (46)
Decreased	4 (0)	115 (3)	241 (6)	360 (9)
Total	541(14)	2,674 (70)	588 (15)	3,803 (100)

D

-nitrogen 3' UTR crosslinking sites (%)

mRNA	*Decreased*	*No change*	*Increased*	*Total*
Increased	99 (3)	1,709 (43)	171 (4)	1,979 (50)
No change	97 (2)	1,642 (42)	160 (4)	1,899 (48)
Decreased	2 (0)	46 (1)	6 (0)	54 (1)
Total	198 (5)	3,397 (86)	337 (9)	3,932

Figure 5.15 Assessment of crosslinking site and mRNA changes in starvation conditions.

(A-B) Enriched GO terms of genes up- and down-regulated upon glucose (A) or nitrogen (B) starvation. (C) The number and percentage of 3' UTR crosslinking sites with indicated changes in crosslinking site coverage and corresponding mRNA expression upon glucose starvation. (D) The number and percentage of 3' UTR crosslinking sites with indicated changes in crosslinking site coverage and corresponding mRNA expression upon nitrogen starvation.

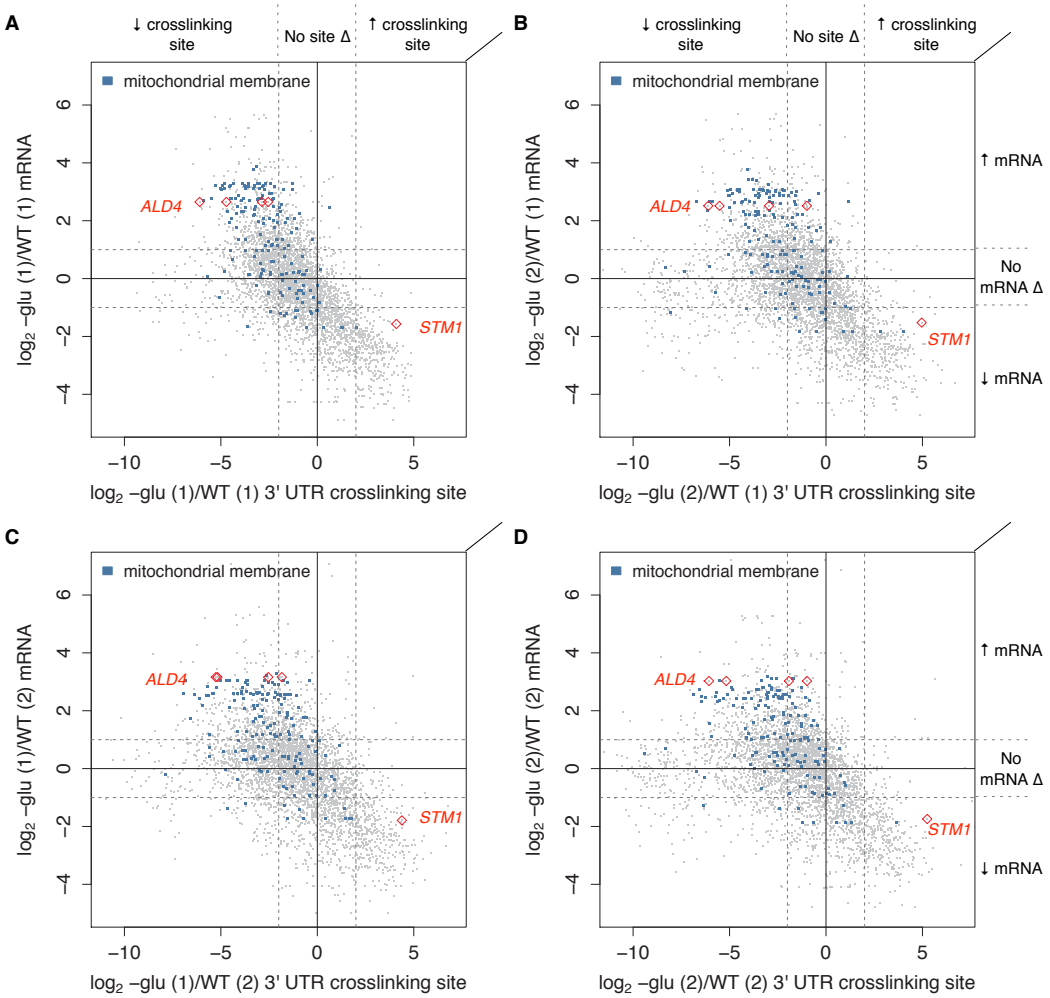


Figure 5.16 Global changes in 3' UTR crosslinking site upon glucose starvation.
Changes in crosslinking site coverage from one replicate library each of WT and glucose starvation conditions are plotted versus changes in the corresponding mRNA from one replicate library each of WT and glucose starvation conditions. Dotted lines and colors are as in Figure 5.6A.

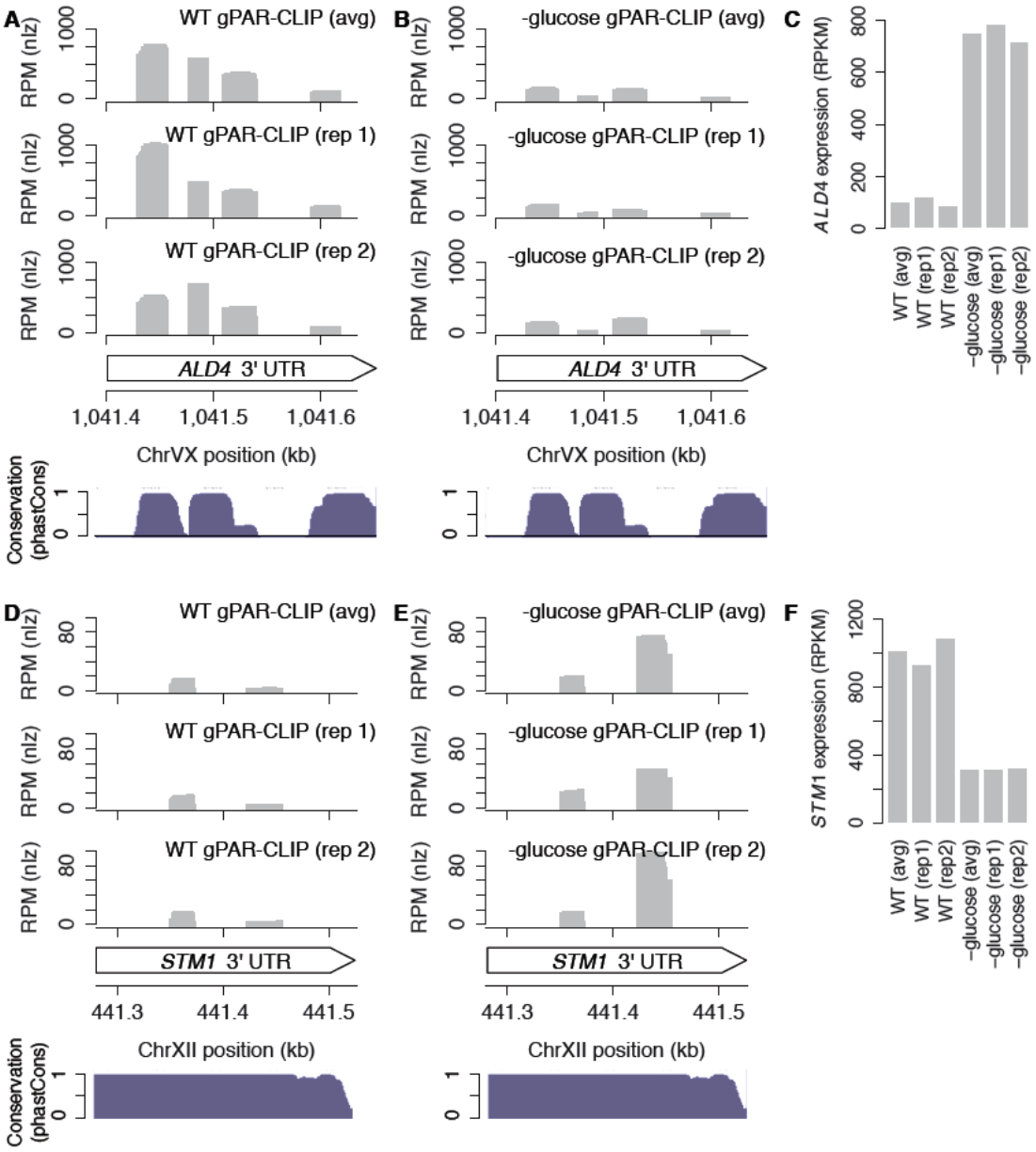


Figure 5.17 Changes in 3' UTR crosslinking sites on *ALD4* and *STM1* upon glucose starvation.
Same as Figure 5.6B-E but showing crosslinking site coverage and mRNA expression in individual replicate libraries.

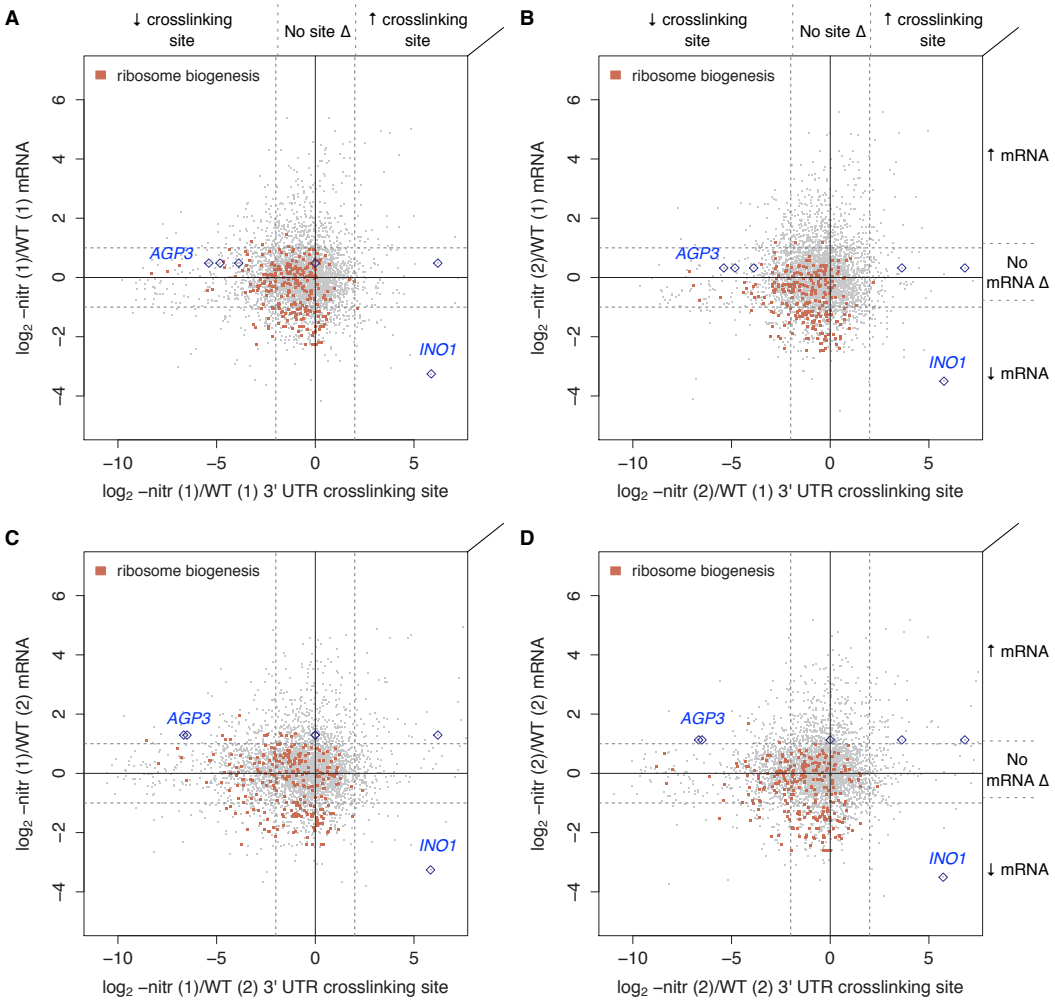


Figure 5.18 Global changes in 3' UTR crosslinking site upon nitrogen starvation.

Changes in crosslinking site coverage from one replicate library each of WT and nitrogen starvation conditions are plotted versus changes in the corresponding mRNA from one replicate library each of WT and nitrogen starvation conditions. Dotted lines and colors are as in Figure 5.7A.

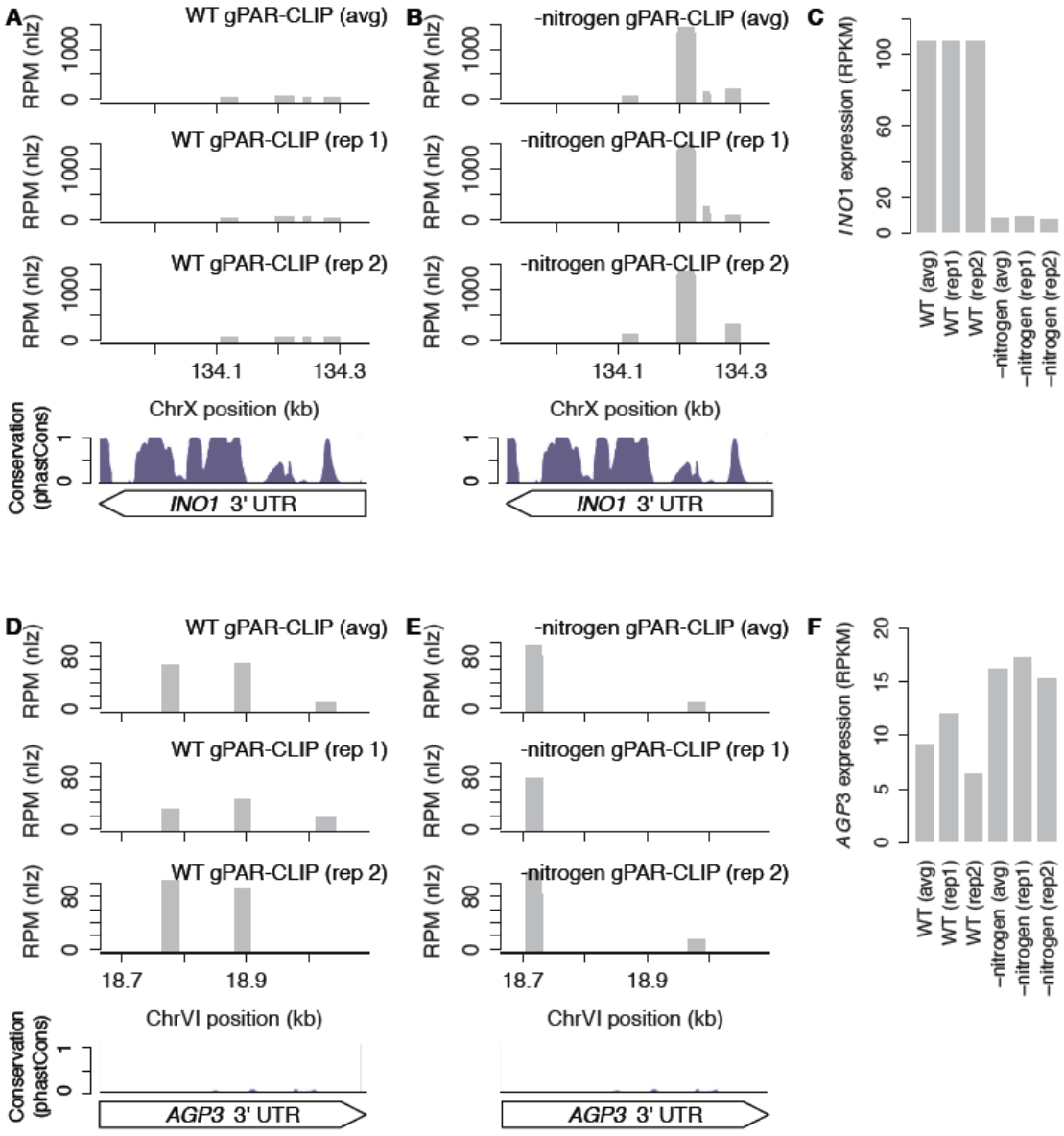


Figure 5.19 Changes in 3' UTR crosslinking sites on *INO1* and *AGP3* upon glucose starvation.
Same as Figure 5.7B-E but showing crosslinking site coverage and mRNA expression in individual replicate libraries.

Chapter 6

Significance and implications

Small RNAs and RNA-binding proteins have emerged as central regulators of development, and have been implicated in a wide range of biological processes, including the maintenance of stem cell pluripotency and genome integrity, the suppression of oncogenesis, and even complex processes like learning and memory. In this thesis, I describe several aspects of small RNA- and RBP- based post-transcriptional regulatory mechanisms in *C. elegans* and *S. cerevisiae*. These findings have provided important insights into analogous mechanisms in higher eukaryotic organisms. In this chapter, I will summarize the key implications in each chapter and discuss future directions.

6.1 26G RNAs and evolution of new genes

Chapter 2 describes the identification and characterization of two classes of germline-generated 26G endo-siRNAs. Class I 26G RNAs are produced during spermatogenesis and restricted to sperm. They target mRNAs expressed during spermatogenesis and loss of class I 26G RNAs is associated with sperm defects. Class II 26G RNAs are produced in oocytes and inherited by the next generation. Targets of class II 26G RNAs are expressed throughout development

and depleted of germline genes. Mutants defective in class II 26G RNAs display wild-type fertility and do not have any obvious developmental phenotypes.

The lack of obvious developmental functions of class II 26G RNAs prompted an evolutionary analysis of their target genes. By comparison with *C. briggsae*, a nematode that diverged from *C. elegans* ~100 million years ago, we and others found that a significantly smaller fraction of Class II 26G RNAs are conserved in *C. briggsae*, suggesting that Class II 26G RNA targets are fast evolving genes (249). In addition, class II 26G RNAs often reside in clusters of duplicated genes (45, 250). Based on these two lines of evidence, we speculate that class II 26G RNA targets arise from gene duplications and are fast evolving. These genes carry with them an anti-sense control mechanism (26G RNA) to curb their own expression. We further speculate that such a mechanism could allow for the rapid emergence of new genes without the danger of overexpressing deleterious gene products. Similar molecular arms races between small RNAs and gene duplications might be present in human genomes as well.

Since the central components of the 26G RNA biogenesis pathway have been identified, it is now possible to understand additional mechanistic details of the pathway using a combination of genetic and biochemical approaches. For example, in Chapter 2, I present genetic evidence that 26G RNAs biogenesis requires Dicer. However, *in vitro* Dicer processing experiments with dsRNA substrates in the past showed that Dicer products are usually 20-24nt in length, shorter than 26nt (93). This apparent discrepancy was resolved by a recent *in*

vitro study suggesting that 26G RNAs can be generated by Dicer processing of dsRNAs with blunt ends (251). By incubating terminally radiolabeled blunt-ended dsRNAs with wild-type embryonic extract, a prominent 26nt RNA species can be observed (Figure 6.1). Using this *in vitro* assay, I examined 26G RNA processing activity in several mutants defective in 26G RNA biogenesis. I found that 26G dicing is compromised in *eri-1*, *dcr-1*, *rrf-3* and *rde-4* mutant extracts, consistent with the model that *DCR-1* functions together in a complex with *ERI-1*, *RRF-3*, and *RDE-4* (Figure 6.1). The fact that *ergo-1* is not required for efficient 26G RNA dicing is consistent with a role of ERGO-1 downstream of Dicer processing (Figure 6.1). Curiously, in all reactions, a >26 nt RNA species can be observed, which accumulates to higher levels in *rde-4* and could represent an intermediate form of the 26G RNA precursor (Figure 6.1). The identity of this putative intermediate and its *in vivo* relevance remains to be studied.

6.2 CK2 substrates and post-translational modifications of miRISC

Chapter 3 describes the identification of a conserved protein kinase, casein kinase 2 (CK2), which genetically and physically interacts with miRISC to control the activity of miRNAs. One immediate follow-up study would be to identify the substrates that mediate CK2's function in the miRNA pathway. We have evidence that one of the key miRISC components, the DEAD box helicase CGH-1, physically interacts with CK2 (Figure 6.2A). Additionally, CK2 can phosphorylate one of the potential CK2 sites on CGH-1 *in vitro* (Figure 6.2B). We are currently testing if CK2 directly phosphorylates CGH-1 *in vivo*, and whether

this phosphorylation event is required for miRNA activity. Another potential CK2 substrate is VIG-1, a RNA-binding protein in miRISC. VIG-1 can also be phosphorylated by CK2 *in vitro* (Figure 6.3). We will test if phosphorylation of VIG-1 by CK2 regulates miRNA activity in the future.

The connection of CK2 to the miRNA pathway reveals the importance of post-translational modifications of miRISC components in controlling gene regulation by miRNAs. Although post-translational modification of proteins is well established as a mechanism of regulating enzymatic activity, protein stability, and signal transduction, its potential role in the miRNA pathway has remained largely unexplored. A potentially fruitful line of investigation would be to systematically perform large-scale mass spec analyses of purified miRISC components to identify additional post-translational modifications and study their roles in controlling miRISC activity. For example, our preliminary mass spec analyses have identified prolyl hydroxylation on AIN-1, a conserved GW182 protein in miRISC, which could control miRISC stability and/or activity. These studies are expected to lead to improved understanding of how miRNAs control gene expression and promises to motivate the development of new therapies for the rapidly expanding list of human diseases whose pathogeneses involve microRNA dysregulation.

6.3 Developmental and tissue-specific 3'UTR isoforms

MiRNAs primarily bind to the 3' untranslated regions (3' UTRs) of target mRNAs and repress their expression. Although thousands of miRNAs have been

identified, the annotations of 3'UTRs are far from complete. Chapter 4 provides a comprehensive map of 3'UTR isoforms for the majority of *C. elegans* genes. This study has quintupled the number of distinct 3'UTR annotations for *C. elegans*, providing empirical evidence for 3'UTR regulatory elements predicted to be targeted by miRNAs. This study was based on the 454 platform for its longer read length (~250nt) to allow unambiguous assignment of 3'UTRs to genes. Over the past few years, the read length capacity of the Illumina sequencing platform has increased dramatically (from ~30nt to ~200nt). We are now shifting to the Illumina Hiseq2000 platform to perform paired-end sequencing of 3'UTR libraries with much higher throughput.

Because many genes are expressed in a highly tissue-specific manner, we are performing tissue- and cell type-specific polyA-captured libraries to obtain a more complete characterization of the C. elegans 3'UTRome. In addition, these studies will allow us to determine if there are tissue-specific biases in the selection and expression of particular 3'UTR isoforms. To this end, we have generated strains with tissue-specific expression of epitope-tagged PolyA Binding Protein (PABP), enabling immunopurification of polyadenylated transcripts from specific tissues (Figure 6.4). These tissue-specific UTR datasets will be integrated with ALG-1 HTS-CLIP datasets also currently in progress in the lab to generate a tissue-specific atlas of the miRNA-mRNA interaction network, allowing better understanding of miRNA-mediated post-transcriptional regulation of gene expression underlying development and differentiation.

6.4 RBPome in yeast and worms

Chapter 5 describes a comprehensive map of RBP crosslinking sites across the budding yeast non-translating mRNA transcriptome and, for the first time, describes the dynamics of mRNA-RBP binding under normal and nutrient-limited growth conditions. Delineating *in vivo* sites of RBP binding will aid in directing future studies for identification of sites responsive to environmental or genetic perturbations, refinement of primary sequence and secondary structural elements recognized by specific RBPs, and elucidation of the complex network of regulatory processes that contribute to regulation of expression of each individual mRNA.

The insights yielded by the gPAR-CLIP approach prompted us to combine in vivo UV crosslinking with mass spectrometry to survey the global RNA-binding proteome. Large-scale *in vitro* screens for RBPs have been performed previously by hybridizing RNAs to microarrays spotted with purified proteins, with the caveats of variation in protein purification and lack of cellular context (252, 253). We reasoned that UV crosslinking could stabilize physiologically relevant RBP-RNA interactions and allow stringent purification steps so that non-specific proteins or proteins indirectly associated with RNAs via other RBPs could be removed. As in gPAR-CLIP, we relied on 4sU to enhance crosslinking efficiency and enriched for mRNA-binding proteins by ribosome depletion and oligo(dT) selection. We took advantage of the large mobility shift of proteins induced by crosslinking to RNA (1 kb RNA ≈ 340 kD) to separate RBPs from free proteins by denaturing SDS-PAGE. Gel pieces containing >330 kD RBP-RNA conjugates

were excised from the gel and subjected to in-gel trypsin digestion. Peptides were identified by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) (Figure 6.5A). Altogether, we identified 259 putative RBPs, out of which 132 contain no recognizable RNA-binding domains (Figure 6.5B). A large number of these newly discovered RBPs are conserved and fulfill roles in critical biological processes including trafficking, cytoskeletal organization, metabolism, and prion biology. Collectively, these results expand the repertoire of post-transcriptional regulatory mechanisms. I am currently analyzing the biological function of several of these novel RBPs and the RNA substrates for several dozen of the noncanonical RBPs by *in vivo* cross-linking and deep sequencing.

gPAR-CLIP is readily applicable to other organisms. I have generated reagents to profile global RNA-protein interactions underlying post-transcriptional regulation in *C. elegans* somatic and germline development. Taking advantage of the *glp-4(bn2ts)* mutant that fails to develop a germline at non-permissive temperatures, I captured the RNA-binding proteins from the somatic tissues of worms, and compared them with those captured from wildtype animals (containing both soma and germline). In our pilot experiments, ~1400 RBPs were identified, with ~400 showing germline enrichment. In the future, we will identify both known and novel RBPs in the datasets and study the functions of interesting candidates.

The identification of such a large number of proteins that lack canonical RNA binding domains in both yeast and worms reveals a profound deficit in our

previous understanding of what constitutes an RBP. Future studies motivated by my findings will likely reveal additional unexpected features of RBP regulation that are universal in all organisms.

6.5 Summary

Small RNAs and RBPs post-transcriptionally regulate RNA stability, localization, and translation, yet understanding of their functions is quite rudimentary in comparison to the proteins that regulate transcription. My thesis studies have revealed new modes of post-transcriptional regulation in model organisms, which are functioning in higher organisms as well.

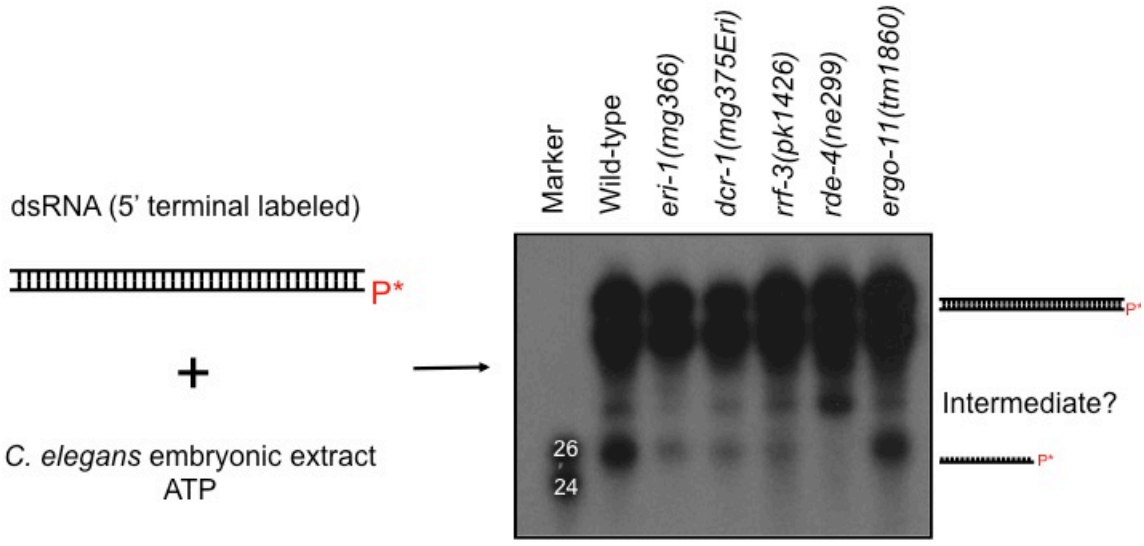


Figure 6.1 Processing of blunt-ended dsRNA *in vitro* with embryonic extracts from wild-type and genetic mutants.
dsRNAs with blunt ends and radiolabeled 5' termini were incubated with embryonic extracts of indicated genotypes. Wild-type extract generates 26nt RNA. *eri-1*, *dcr-1*, *rrf-3*, and *rde-4* mutant extracts exhibit compromised *Dicer* activity, while *ergo-1* mutant extract shows normal processing activity. A >26nt RNA species is also observed, which accumulates to higher levels in *rde-4* relative to wild-type.

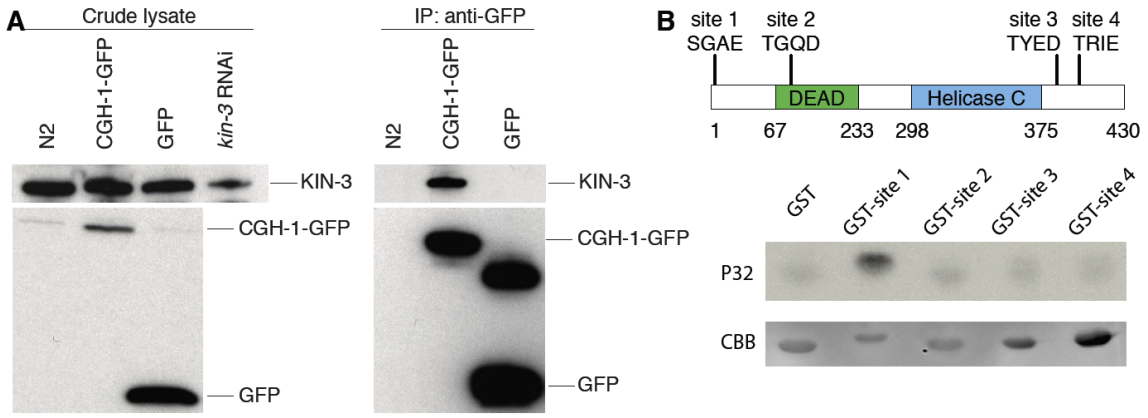


Figure 6.2 KIN-3 co-immunoprecipitates (co-IP) with CGH-1 and phosphorylates CGH-1 *in vitro*.
(A) CGH-1-GFP was immunoprecipitated with monoclonal anti-GFP antibody from worms expressing CGH-1-GFP or from wild-type worms (as negative control). CGH-1-GFP immunoprecipitates (IP) were blotted for both KIN-3 and GFP. KIN-3 is present in the CGH-1-GFP IP. (B) Peptide fragments (20 aa in length) flanking each of the four putative CK2 phosphorylation sites on CGH-1 were fused to GST, expressed in *E. coli*, purified, and subjected to CK2 *in vitro* phosphorylation. Only the site 1 fusion peptide can be phosphorylated by CK2 *in vitro*.

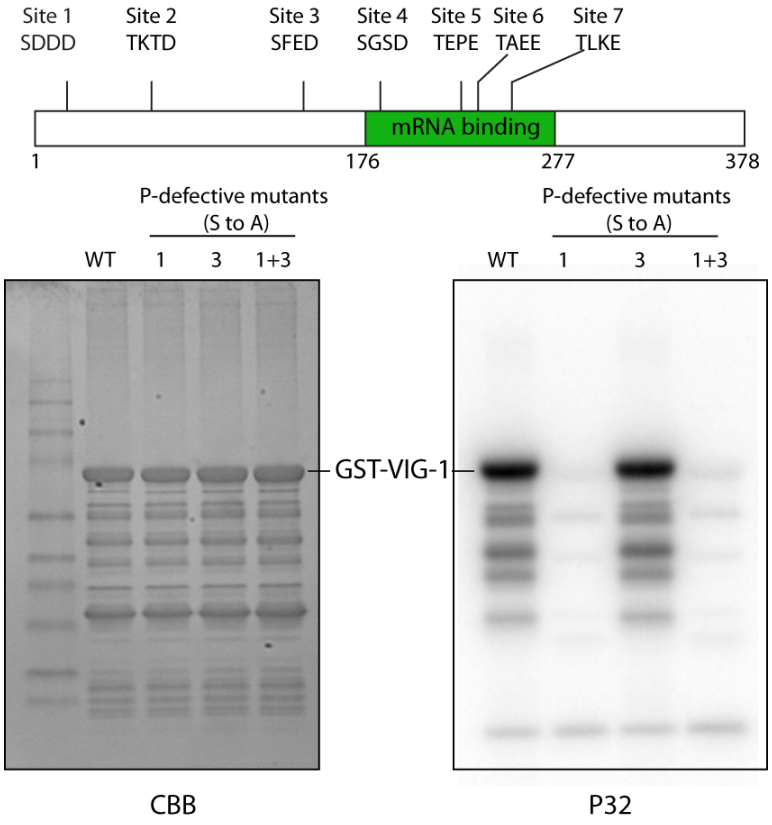


Figure 6.3 CK2 phosphorylates VIG-1 *in vitro*.
Wildtype and phosphodefective forms of VIG-1 were fused to GST, expressed and purified from *E. coli*, and subjected to *in vitro* phosphorylation by CK2. Wildtype VIG-1 can be phosphorylated by CK2, but site 1 phosphodefective mutation (serine to alanine) on VIG-1 specifically blocks CK2 phosphorylation.

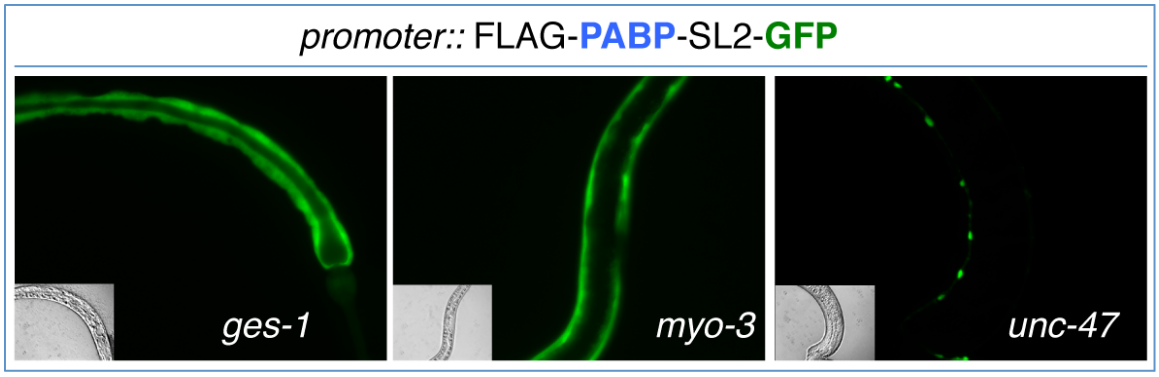


Figure 6.4 PABP strains for profiling tissue-specific 3'UTR isoforms.
Transgenic strains expressing FLAG-PABP-SL2-GFP transgene in three major tissue types of *C. elegans*: intestine (*ges-1*::PABP), body wall muscle (*myo-3*::PABP), and GABAergic neurons (*unc-47*::PABP). Image courtesy: Vishal Khivansara.

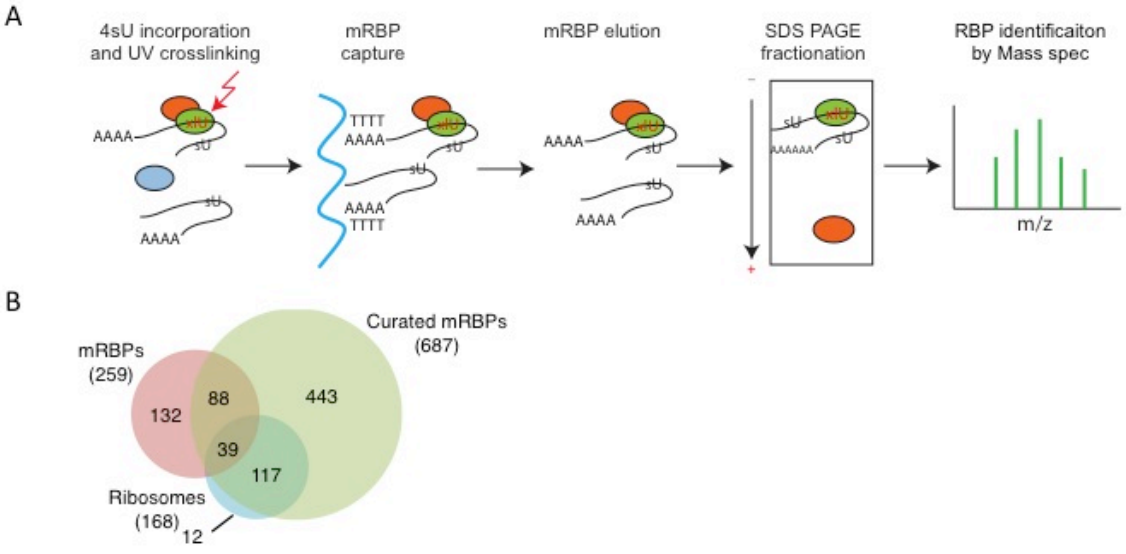


Figure 6.5 Survey of the yeast RNA-binding proteome by mass spectrometry.
(a) Schematic of RBP identification by mass spectrometry. (B) Comparison of RBPs identified by mass spec and a list of bioinformatically curated RBPs and ribosome proteins.

BIBLIOGRAPHY

1. Ingolia NT, Ghaemmaghami S, Newman JR, & Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324(5924):218-223.
2. Wang Y*, et al.* (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* 99(9):5860-5865.
3. Lecuyer E*, et al.* (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131(1):174-187.
4. Sonenberg N & Hinnebusch AG (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* 136(4):731-745.
5. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116(2):281-297.
6. Dreyfuss G, Kim VN, & Kataoka N (2002) Messenger-RNA-binding proteins and the messages they carry. *Nature reviews. Molecular cell biology* 3(3):195-205.
7. Jackson JS, Jr., Houshmandi SS, Lopez Leban F, & Olivas WM (2004) Recruitment of the Puf3 protein to its mRNA target for regulation of mRNA decay in yeast. *RNA* 10(10):1625-1636.
8. Kim VN, Han J, & Siomi MC (2009) Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* 10(2):126-139.
9. Chen X (2009) Small RNAs and their roles in plant development. *Annual review of cell and developmental biology* 25:21-44.
10. Thomson T & Lin H (2009) The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annual review of cell and developmental biology* 25:355-376.
11. Okamura K & Lai EC (2008) Endogenous small interfering RNAs in animals. *Nat Rev Mol Cell Biol* 9(9):673-678.
12. Ghildiyal M*, et al.* (2008) Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science* 320(5879):1077-1081.
13. Gu W*, et al.* (2009) Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the C. elegans germline. *Molecular cell* 36(2):231-244.
14. Guang S*, et al.* (2008) An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* 321(5888):537-541.
15. Lee RC, Feinbaum RL, & Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75(5):843-854.
16. Reinhart BJ*, et al.* (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 403(6772):901-906.

17. Kim VN (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* 6(5):376-385.
18. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136(2):215-233.
19. Bagga S*, et al.* (2005) Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* 122(4):553-563.
20. Filipowicz W, Bhattacharyya SN, & Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* 9(2):102-114.
21. Giraldez AJ*, et al.* (2006) Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* 312(5770):75-79.
22. Kiriakidou M*, et al.* (2007) An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* 129(6):1141-1151.
23. Mathonnet G*, et al.* (2007) MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science* 317(5845):1764-1767.
24. Bazzini AA, Lee MT, & Giraldez AJ (Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* 336(6078):233-237.
25. Djuranovic S, Nahvi A, & Green R (miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science* 336(6078):237-240.
26. Siomi MC, Sato K, Pezic D, & Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12(4):246-258.
27. Saito K*, et al.* (2006) Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. *Genes & development* 20(16):2214-2222.
28. Vagin VV*, et al.* (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* 313(5785):320-324.
29. Aravin A*, et al.* (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442(7099):203-207.
30. Girard A, Sachidanandam R, Hannon GJ, & Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442(7099):199-202.
31. Lau NC, Lim LP, Weinstein EG, & Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* 294(5543):858-862.
32. Ishizu H, Siomi H, & Siomi MC (2012) Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Genes Dev* 26(21):2361-2373.
33. Brennecke J*, et al.* (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* 128(6):1089-1103.
34. Gunawardane LS*, et al.* (2007) A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science* 315(5818):1587-1590.

35. Kawaoka S, Izumi N, Katsuma S, & Tomari Y (2011) 3' end formation of PIWI-interacting RNAs in vitro. *Mol Cell* 43(6):1015-1022.
36. Grimson A*, et al.* (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455(7217):1193-1197.
37. Ruby JG*, et al.* (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* 127(6):1193-1207.
38. Batista PJ*, et al.* (2008) PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Molecular cell* 31(1):67-78.
39. Das PP*, et al.* (2008) Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline. *Mol Cell* 31(1):79-90.
40. Bagijn MP*, et al.* (2012) Function, Targets, and Evolution of Caenorhabditis elegans piRNAs. *Science*.
41. Lee HC*, et al.* (2012) C. elegans piRNAs Mediate the Genome-wide Surveillance of Germline Transcripts. *Cell*.
42. Cecere G, Zheng GX, Mansisidor AR, Klymko KE, & Grishok A (2012) Promoters recognized by forkhead proteins exist for individual 21U-RNAs. *Mol Cell* 47(5):734-745.
43. Gu W*, et al.* (2012) CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. *Cell* 151(7):1488-1500.
44. Han T*, et al.* (2009) 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in Caenorhabditis elegans. *Proc Natl Acad Sci USA* 106(44):18674-18679.
45. Vasale JJ*, et al.* (2010) Sequential rounds of RNA-dependent RNA transcription drive endogenous small-RNA biogenesis in the ERGO-1/Argonaute pathway. *Proceedings of the National Academy of Sciences of the United States of America* 107(8):3582-3587.
46. Conine CC*, et al.* (2010) Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in Caenorhabditis elegans. *Proc Natl Acad Sci USA* 107(8):3588-3593.
47. Claycomb JM*, et al.* (2009) The Argonaute CSR-1 and its 22G-RNA cofactors are required for holocentric chromosome segregation. *Cell* 139(1):123-134.
48. Ashe A*, et al.* (2012) piRNAs can trigger a multigenerational epigenetic memory in the germline of C. elegans. *Cell* 150(1):88-99.
49. Buckley BA*, et al.* (2012) A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature* 489(7416):447-451.
50. Shirayama M*, et al.* (2012) piRNAs initiate an epigenetic memory of nonself RNA in the C. elegans germline. *Cell* 150(1):65-77.
51. Burton NO, Burkhart KB, & Kennedy S (2011) Nuclear RNAi maintains heritable gene silencing in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* 108(49):19683-19688.

52. Avgousti DC, Palani S, Sherman Y, & Grishok A (2012) CSR-1 RNAi pathway positively regulates histone expression in C. elegans. *EMBO J* 31(19):3821-3832.
53. Das PP*, et al.* (2008) Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline. *Molecular cell* 31(1):79-90.
54. Martin KC & Ephrussi A (2009) mRNA localization: gene expression in the spatial dimension. *Cell* 136(4):719-730.
55. Moore MJ & Proudfoot NJ (2009) Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* 136(4):688-700.
56. Glisovic T, Bachorik JL, Yong J, & Dreyfuss G (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 582(14):1977-1986.
57. Hafidh S, Capkova V, & Honys D (2011) Safe keeping the message: mRNP complexes tweaking after transcription. *Adv Exp Med Biol* 722:118-136.
58. Anderson P & Kedersha N (2009) RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nature reviews. Molecular cell biology* 10(6):430-436.
59. Moore MJ (2005) From birth to death: the complex lives of eukaryotic mRNAs. *Science* 309(5740):1514-1518.
60. Bohl F, Kruse C, Frank A, Ferring D, & Jansen RP (2000) She2p, a novel RNA-binding protein tethers ASH1 mRNA to the Myo4p myosin motor via She3p. *EMBO J* 19(20):5514-5524.
61. Kato M*, et al.* (2012) Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* 149(4):753-767.
62. Han TW*, et al.* (2012) Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell* 149(4):768-779.
63. Jensen KB, Musunuru K, Lewis HA, Burley SK, & Darnell RB (2000) The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proceedings of the National Academy of Sciences of the United States of America* 97(11):5740-5745.
64. Wei WJ*, et al.* (2012) YB-1 binds to CAUC motifs and stimulates exon inclusion by enhancing the recruitment of U2AF to weak polypyrimidine tracts. *Nucleic acids research*.
65. Tenenbaum SA, Carson CC, Lager PJ, & Keene JD (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proceedings of the National Academy of Sciences of the United States of America* 97(26):14085-14090.
66. Gerber AP, Herschlag D, & Brown PO (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS biology* 2(3):E79.
67. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, & Brown PO (2008) Diverse RNA-binding proteins interact with functionally related sets of

RNAs, suggesting an extensive regulatory system. *PLoS biology* 6(10):e255.
68. Riordan DP, Herschlag D, & Brown PO (2010) Identification of RNA recognition elements in the Saccharomyces cerevisiae transcriptome. *Nucleic acids research* 39(4):1501-1509.
69. Mili S & Steitz JA (2004) Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA* 10(11):1692-1694.
70. Darnell JC, Mostovetsky O, & Darnell RB (2005) FMRP RNA targets: identification and validation. *Genes Brain Behav* 4(6):341-349.
71. Licatalosi DD*, et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456(7221):464-469.
72. Ule J*, et al.* (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302(5648):1212-1215.
73. Ule J, Jensen K, Mele A, & Darnell RB (2005) CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* 37(4):376-386.
74. Hafner M*, et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141(1):129-141.
75. Chi SW, Zang JB, Mele A, & Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460(7254):479-486.
76. Zisoulis DG*, et al.* (2010) Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nature structural & molecular biology* 17(2):173-179.
77. Mukherjee N*, et al.* (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Molecular cell* 43(3):327-339.
78. Lebedeva S*, et al.* (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Molecular cell* 43(3):340-352.
79. Creamer TJ*, et al.* (2011) Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genet* 7(10):e1002329.
80. Castello A*, et al.* (2012) Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* 149(6):1393-1406.
81. Baltz AG*, et al.* (2012) The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular cell* 46(5):674-690.
82. Aravin AA, Hannon GJ, & Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318(5851):761-764.
83. Ambros V, Lee RC, Lavanway A, Williams PT, & Jewell D (2003) MicroRNAs and other tiny endogenous RNAs in C. elegans. *Curr Biol* 13(10):807-818.
84. Duchaine TF*, et al.* (2006) Functional proteomics reveals the biochemical niche of C. elegans DCR-1 in multiple small-RNA-mediated pathways. *Cell* 124(2):343-354.

85. Lee RC, Hammell CM, & Ambros V (2006) Interacting endogenous and exogenous RNAi pathways in Caenorhabditis elegans. *RNA* 12(4):589-597.
86. Watanabe T*, et al.* (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453(7194):539-543.
87. Tam OH*, et al.* (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453(7194):534-538.
88. Okamura K*, et al.* (2008) The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* 453(7196):803-806.
89. Kawamura Y*, et al.* (2008) Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature* 453(7196):793-797.
90. Czech B*, et al.* (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature* 453(7196):798-802.
91. Cox DN*, et al.* (1998) A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev* 12(23):3715-3727.
92. Knight SW & Bass BL (2001) A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science* 293(5538):2269-2271.
93. Ketting RF*, et al.* (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev* 15(20):2654-2659.
94. Grishok A*, et al.* (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell* 106(1):23-34.
95. Wang G & Reinke V (2008) A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. *Curr Biol* 18(12):861-867.
96. Batista PJ*, et al.* (2008) PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Mol Cell* 31(1):67-78.
97. Brennecke J*, et al.* (2008) An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* 322(5906):1387-1392.
98. Lau NC*, et al.* (2006) Characterization of the piRNA complex from rat testes. *Science* 313(5785):363-367.
99. Beanan MJ & Strome S (1992) Characterization of a germ-line proliferation mutation in C. elegans. *Development* 116(3):755-766.
100. Kennedy S, Wang D, & Ruvkun G (2004) A conserved siRNA-degrading RNase negatively regulates RNA interference in C. elegans. *Nature* 427(6975):645-649.
101. Reinke V, Gil IS, Ward S, & Kazmer K (2004) Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans. *Development* 131(2):311-323.
102. Asikainen S, Storvik M, Lakso M, & Wong G (2007) Whole genome microarray analysis of C. elegans rrf-3 and eri-1 mutants. *FEBS Lett* 581(26):5050-5054.

103. Makeyev EV & Bamford DH (2002) Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes. *Mol Cell* 10(6):1417-1427.
104. Sijen T*, et al.* (2001) On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* 107(4):465-476.
105. Smardon A*, et al.* (2000) EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in C. elegans. *Curr Biol* 10(4):169-178.
106. Faehnle CR & Joshua-Tor L (2007) Argonautes confront new small RNAs. *Curr Opin Chem Biol* 11(5):569-577.
107. Yigit E*, et al.* (2006) Analysis of the C. elegans Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* 127(4):747-757.
108. Liu J*, et al.* (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science* 305(5689):1437-1441.
109. Simmer F*, et al.* (2002) Loss of the putative RNA-directed RNA polymerase RRF-3 makes C. elegans hypersensitive to RNAi. *Curr Biol* 12(15):1317-1319.
110. Pak J & Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science* 315(5809):241-244.
111. Sijen T, Steiner FA, Thijssen KL, & Plasterk RH (2007) Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* 315(5809):244-247.
112. L'Hernault SW & Roberts TM (1995) Cell biology of nematode sperm. *Methods Cell Biol* 48:273-301.
113. Lu C, Meyers BC, & Green PJ (2007) Construction of small RNA cDNA libraries for deep sequencing. *Methods* 43(2):110-117.
114. Pall GS & Hamilton AJ (2008) Improved northern blot method for enhanced detection of small RNA. *Nat Protoc* 3(6):1077-1084.
115. Lee MH & Schedl T (2006) RNA in situ hybridization of dissected gonads. *WormBook*:1-7.
116. Kamath RS & Ahringer J (2003) Genome-wide RNAi screening in Caenorhabditis elegans. *Methods* 30(4):313-321.
117. Spike CA, Bader J, Reinke V, & Strome S (2008) DEPS-1 promotes P-granule assembly and RNA interference in C. elegans germ cells. *Development* 135(5):983-993.
118. Hodgkin J, Papp A, Pulak R, Ambros V, & Anderson P (1989) A new kind of informational suppression in the nematode Caenorhabditis elegans. *Genetics* 123(2):301-313.
119. Aoki K, Moriguchi H, Yoshioka T, Okawa K, & Tabara H (2007) In vitro analyses of the production and activity of secondary small interfering RNAs in C. elegans. *EMBO J* 26(24):5007-5019.
120. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403-410.
121. Li C & Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* 98(1):31-36.

122. Brenner S, Hodgkin,J , Horvitz, R . (1979) Nondisjunction mutants of the nematode C. elegans. *Genetics* (91):67-94.
123. Ward S, Argon Y, & Nelson GA (1981) Sperm morphogenesis in wild-type and fertilization-defective mutants of Caenorhabditis elegans. *J Cell Biol* 91(1):26-44.
124. Stiernagle T (2006) Maintenance of C. elegans. *WormBook*:1-11.
125. Chen C*, et al.* (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res* 33(20):e179.
126. Nolan T, Hands RE, & Bustin SA (2006) Quantification of mRNA using real-time RT-PCR. *Nat Protoc* 1(3):1559-1582.
127. Tops BB, Plasterk RH, & Ketting RF (2006) The Caenorhabditis elegans Argonautes ALG-1 and ALG-2: almost identical yet different. *Cold Spring Harb Symp Quant Biol* 71:189-194.
128. Ding L, Spencer A, Morita K, & Han M (2005) The developmental timing regulator AIN-1 interacts with miRISCs and may target the argonaute protein ALG-1 to cytoplasmic P bodies in C. elegans. *Mol Cell* 19(4):437-447.
129. Zhang L*, et al.* (2007) Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol Cell* 28(4):598-613.
130. Hammell CM, Lubin I, Boag PR, Blackwell TK, & Ambros V (2009) nhl-2 Modulates microRNA activity in Caenorhabditis elegans. *Cell* 136(5):926-938.
131. Caudy AA*, et al.* (2003) A micrococcal nuclease homologue in RNAi effector complexes. *Nature* 425(6956):411-414.
132. Parry DH, Xu J, & Ruvkun G (2007) A whole-genome RNAi Screen for C. elegans miRNA pathway genes. *Curr Biol* 17(23):2013-2022.
133. Kim JK*, et al.* (2005) Functional genomic analysis of RNA interference in C. elegans. *Science* 308(5725):1164-1167.
134. Dominguez I, Sonenshein GE, & Seldin DC (2009) Protein kinase CK2 in health and disease: CK2 and its role in Wnt and NF-kappaB signaling: linking development and cancer. *Cell Mol Life Sci* 66(11-12):1850-1857.
135. St-Denis NA & Litchfield DW (2009) Protein kinase CK2 in health and disease: From birth to death: the role of protein kinase CK2 in the regulation of cell proliferation and survival. *Cell Mol Life Sci* 66(11-12):1817-1829.
136. Trembley JH, Wang G, Unger G, Slaton J, & Ahmed K (2009) Protein kinase CK2 in health and disease: CK2: a key player in cancer biology. *Cell Mol Life Sci* 66(11-12):1858-1867.
137. Niefind K, Raaf J, & Issinger OG (2009) Protein kinase CK2 in health and disease: Protein kinase CK2: from structures to insights. *Cell Mol Life Sci* 66(11-12):1800-1816.
138. Hu E & Rubin CS (1990) Casein kinase II from Caenorhabditis elegans. Properties and developmental regulation of the enzyme; cloning and sequence analyses of cDNA and the gene for the catalytic subunit. *J Biol Chem* 265(9):5072-5080.

139. Li M, Jones-Rhoades MW, Lau NC, Bartel DP, & Rougvie AE (2005) Regulatory mutations of mir-48, a C. elegans let-7 family MicroRNA, cause developmental timing defects. *Dev Cell* 9(3):415-422.
140. Abbott AL*, et al.* (2005) The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans. *Dev Cell* 9(3):403-414.
141. Hayes GD, Frand AR, & Ruvkun G (2006) The mir-84 and let-7 paralogous microRNA genes of Caenorhabditis elegans direct the cessation of molting via the conserved nuclear hormone receptors NHR-23 and NHR-25. *Development* 133(23):4631-4641.
142. Abrahante JE, Miller EA, & Rougvie AE (1998) Identification of heterochronic mutants in Caenorhabditis elegans. Temporal misexpression of a collagen::green fluorescent protein fusion gene. *Genetics* 149(3):1335-1351.
143. Koh K & Rothman JH (2001) ELT-5 and ELT-6 are required continuously to regulate epidermal seam cell differentiation and cell fusion in C. elegans. *Development* 128(15):2867-2880.
144. Liu Z, Kirch S, & Ambros V (1995) The Caenorhabditis elegans heterochronic gene pathway controls stage-specific transcription of collagen genes. *Development* 121(8):2471-2478.
145. Johnston RJ & Hobert O (2003) A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature* 426(6968):845-849.
146. Schmitz C, Kinge P, & Hutter H (2007) Axon guidance genes identified in a large-scale RNAi screen using the RNAi-hypersensitive Caenorhabditis elegans strain nre-1(hd20) lin-15b(hd126). *Proc Natl Acad Sci U S A* 104(3):834-839.
147. Alvarez-Saavedra E & Horvitz HR (Many families of C. elegans microRNAs are not essential for development or viability. *Curr Biol* 20(4):367-373.
148. Johnson SM*, et al.* (2005) RAS is regulated by the let-7 microRNA family. *Cell* 120(5):635-647.
149. Sternberg PW (2005) Vulval development. *WormBook*:1-28.
150. Eisenmann DM & Kim SK (1997) Mechanism of activation of the Caenorhabditis elegans ras homologue let-60 by a novel, temperature-sensitive, gain-of-function mutation. *Genetics* 146(2):553-565.
151. Ding XC & Grosshans H (2009) Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. *EMBO J* 28(3):213-222.
152. Slack FJ*, et al.* (2000) The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Mol Cell* 5(4):659-669.
153. Hammell CM, Karp X, & Ambros V (2009) A feedback circuit involving let-7-family miRNAs and DAF-12 integrates environmental signals and developmental timing in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* 106(44):18668-18673.

154. Grosshans H, Johnson T, Reinert KL, Gerstein M, & Slack FJ (2005) The temporal patterning microRNA let-7 regulates several transcription factors at the larval to adult transition in C. elegans. *Dev Cell* 8(3):321-330.
155. Wightman B, Ha I, & Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* 75(5):855-862.
156. Chan SP & Slack FJ (2009) Ribosomal protein RPS-14 modulates let-7 microRNA function in Caenorhabditis elegans. *Dev Biol* 334(1):152-160.
157. Simon DJ*, et al.* (2008) The microRNA miR-1 regulates a MEF-2-dependent retrograde signal at neuromuscular junctions. *Cell* 133(5):903-915.
158. Kim YK, Heo I, & Kim VN (Modifications of small RNAs and their associated proteins. *Cell* 143(5):703-709.
159. Qi HH*, et al.* (2008) Prolyl 4-hydroxylation regulates Argonaute 2 stability. *Nature* 455(7211):421-424.
160. Zeng Y, Sankala H, Zhang X, & Graves PR (2008) Phosphorylation of Argonaute 2 at serine-387 facilitates its localization to processing bodies. *Biochem J* 413(3):429-436.
161. Rybak A*, et al.* (2009) The let-7 target gene mouse lin-41 is a stem cell specific E3 ubiquitin ligase for the miRNA pathway protein Ago2. *Nat Cell Biol* 11(12):1411-1420.
162. Poole A*, et al.* (2005) A global view of CK2 function and regulation. *Mol Cell Biochem* 274(1-2):163-170.
163. Olsten ME, Weber JE, & Litchfield DW (2005) CK2 interacting proteins: emerging paradigms for CK2 regulation? *Mol Cell Biochem* 274(1-2):115-124.
164. Zielinska DF, Gnad F, Jedrusik-Bode M, Wisniewski JR, & Mann M (2009) Caenorhabditis elegans has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J Proteome Res* 8(8):4039-4049.
165. Mello CC, Kramer JM, Stinchcomb D, & Ambros V (1991) Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. *EMBO J* 10(12):3959-3970.
166. Fraser AG*, et al.* (2000) Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature* 408(6810):325-330.
167. Rual JF*, et al.* (2004) Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library. *Genome Res* 14(10B):2162-2168.
168. Pall GS, Codony-Servat C, Byrne J, Ritchie L, & Hamilton A (2007) Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. *Nucleic Acids Res* 35(8):e60.
169. Reinhart BJ & Ruvkun G (2001) Isoform-specific mutations in the Caenorhabditis elegans heterochronic gene lin-14 affect stage-specific patterning. *Genetics* 157(1):199-209.

170. Sonoda H, Takamatsu J, & Takahashi S (1995) [Evaluation of 26G pencil point spinal needle in combined epidural-spinal anesthesia]. *Masui. The Japanese journal of anesthesiology* 44(10):1410-1414.
171. Gu W*, et al.* (2009) Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the C. elegans germline. *Mol Cell* 36(2):231-244.
172. de Moor CH, Meijer H, & Lissenden S (2005) Mechanisms of translational control by the 3' UTR in development and differentiation. *Semin Cell Dev Biol* 16(1):49-58.
173. Wickens M, Bernstein DS, Kimble J, & Parker R (2002) A PUF family portrait: 3'UTR regulation as a way of life. *Trends Genet* 18(3):150-157.
174. He L*, et al.* (2005) A microRNA polycistron as a potential human oncogene. *Nature* 435(7043):828-833.
175. Chatterjee S & Pal JK (2009) Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. *Biol Cell* 101(5):251-262.
176. Stein L*, et al.* (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res* 29(1):82-86.
177. Carninci P*, et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309(5740):1559-1563.
178. Thierry-Mieg D & Thierry-Mieg J (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* 7 Suppl 1:S12 11-14.
179. Birney E*, et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799-816.
180. Mitiku N & Baker JC (2007) Genomic analysis of gastrulation and organogenesis in the mouse. *Dev Cell* 13(6):897-907.
181. Cloonan N*, et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 5(7):613-619.
182. Hillier LW*, et al.* (2009) Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res* 19(4):657-666.
183. Tian B, Hu J, Zhang H, & Lutz CS (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* 33(1):201-212.
184. Ji Z, Lee JY, Pan Z, Jiang B, & Tian B (2009) Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A* 106(17):7028-7033.
185. Sandberg R, Neilson JR, Sarma A, Sharp PA, & Burge CB (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 320(5883):1643-1647.
186. Zhang H, Lee JY, & Tian B (2005) Biased alternative polyadenylation in human tissues. *Genome Biol* 6(12):R100.
187. Colgan DF & Manley JL (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev* 11(21):2755-2766.

188. Manley JL (1983) Accurate and specific polyadenylation of mRNA precursors in a soluble whole-cell lysate. *Cell* 33(2):595-605.
189. Murthy KG & Manley JL (1992) Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. *J Biol Chem* 267(21):14804-14811.
190. Dupuy D*, et al.* (2004) A first version of the Caenorhabditis elegans Promoterome. *Genome Res* 14(10B):2169-2175.
191. Reboul J*, et al.* (2003) C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat Genet* 34(1):35-41.
192. Mangone M, Macmenamin P, Zegar C, Piano F, & Gunsalus KC (2008) UTRome.org: a platform for 3'UTR biology in C. elegans. *Nucleic Acids Res* 36(Database issue):D57-62.
193. Shin H*, et al.* (2008) Transcriptome analysis for Caenorhabditis elegans based on novel expressed sequence tags. *BMC Biol* 6:30.
194. Blumenthal T*, et al.* (2002) A global analysis of Caenorhabditis elegans operons. *Nature* 417(6891):851-854.
195. Liu Y, Huang T, MacMorris M, & Blumenthal T (2001) Interplay between AAUAAA and the trans-splice site in processing of a Caenorhabditis elegans operon pre-mRNA. *RNA* 7(2):176-181.
196. Wang ZF, Whitfield ML, Ingledue TC, 3rd, Dominski Z, & Marzluff WF (1996) The protein that binds the 3' end of histone mRNA: a novel RNA-binding protein required for histone pre-mRNA processing. *Genes Dev* 10(23):3028-3040.
197. Marzluff WF, Wagner EJ, & Duronio RJ (2008) Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat Rev Genet* 9(11):843-854.
198. Keall R, Whitelaw S, Pettitt J, & Muller B (2007) Histone gene expression and histone mRNA 3' end structure in Caenorhabditis elegans. *BMC Mol Biol* 8:51.
199. Lall S*, et al.* (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol* 16(5):460-471.
200. Zisoulis DG*, et al.* (2010) Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat Struct Mol Biol* 17(2):173-179.
201. Okamura K, Balla S, Martin R, Liu N, & Lai EC (2008) Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. *Nat Struct Mol Biol* 15(6):581-590.
202. Lund E, Liu M, Hartley RS, Sheets MD, & Dahlberg JE (2009) Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos. *RNA* 15(12):2351-2363.
203. Vaglio P*, et al.* (2003) WorfDB: the Caenorhabditis elegans ORFeome Database. *Nucleic Acids Res* 31(1):237-240.
204. Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, & Sugano S (1997) Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* 200(1-2):149-156.

205. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12(4):656-664.
206. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
207. Lopez R, Silventoinen V, Robinson S, Kibria A, & Gish W (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res* 31(13):3795-3798.
208. Blankenberg D*, et al.* (Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* Chapter 19:Unit 19 10 11-21.
209. Karolchik D*, et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32(Database issue):D493-496.
210. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, & Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 34(Database issue):D140-144.
211. Friedlander MR*, et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26(4):407-415.
212. Krek A*, et al.* (2005) Combinatorial microRNA target predictions. *Nat Genet* 37(5):495-500.
213. Sheets MD, Ogg SC, & Wickens MP (1990) Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res* 18(19):5799-5805.
214. Henikoff S, Kelly JD, & Cohen EH (1983) Transcription terminates in yeast distal to a control sequence. *Cell* 33(2):607-614.
215. Sugimoto Y*, et al.* (2012) Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* 13(8):R67.
216. Olivas W & Parker R (2000) The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. *The EMBO journal* 19(23):6602-6611.
217. Valley CT*, et al.* (2012) Patterns and plasticity in RNA-protein interactions enable recruitment of multiple proteins through a single site. *Proceedings of the National Academy of Sciences of the United States of America* 109(16):6054-6059.
218. Ashe MP, De Long SK, & Sachs AB (2000) Glucose depletion rapidly inhibits translation initiation in yeast. *Molecular biology of the cell* 11(3):833-848.
219. Gallego C, Gari E, Colomina N, Herrero E, & Aldea M (1997) The Cln3 cyclin is down-regulated by translational repression and degradation during the G1 arrest caused by nitrogen deprivation in budding yeast. *The EMBO journal* 16(23):7196-7206.
220. Simpson CE & Ashe MP (2012) Adaptation to stress in yeast: to translate or not? *Biochem Soc Trans* 40(4):794-799.
221. Berglund JA, Chua K, Abovich N, Reed R, & Rosbash M (1997) The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell* 89(5):781-787.

222. Garrey SM, Voelker R, & Berglund JA (2006) An extended RNA binding site for the yeast branch point-binding protein and the role of its zinc knuckle domains in RNA binding. *The Journal of biological chemistry* 281(37):27443-27453.
223. Joo YJ, Kim JH, Kang UB, Yu MH, & Kim J (2011) Gcn4p-mediated transcriptional repression of ribosomal protein genes under amino-acid starvation. *The EMBO journal* 30(5):859-872.
224. Ozsolak F*, et al.* (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143(6):1018-1029.
225. Siepel A*, et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* 15(8):1034-1050.
226. Chartrand P, Meng XH, Singer RH, & Long RM (1999) Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. *Current biology : CB* 9(6):333-336.
227. Gonzalez I, Buonomo SB, Nasmyth K, & von Ahsen U (1999) ASH1 mRNA localization in yeast involves multiple secondary structural elements and Ash1 protein translation. *Current biology : CB* 9(6):337-340.
228. Bernhart SH, Hofacker IL, & Stadler PF (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics* 22(5):614-615.
229. Washietl S, Hofacker IL, & Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America* 102(7):2454-2459.
230. Steigele S, Huber W, Stocsits C, Stadler PF, & Nieselt K (2007) Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions. *BMC Biol* 5:25.
231. Hohmann S & Mager WH (2003) *Yeast stress responses* (Springer, Berlin ; New York) [Rev. Ed pp xiii, 389 p.
232. Larochelle M, Drouin S, Robert F, & Turcotte B (2006) Oxidative stress-activated zinc cluster protein Stb5 has dual activator/repressor functions required for pentose phosphate pathway regulation and NADPH production. *Molecular and cellular biology* 26(17):6690-6701.
233. Vyas VK, Berkey CD, Miyao T, & Carlson M (2005) Repressors Nrg1 and Nrg2 regulate a set of stress-responsive genes in Saccharomyces cerevisiae. *Eukaryot Cell* 4(11):1882-1891.
234. Van Dyke N, Baby J, & Van Dyke MW (2006) Stm1p, a ribosome-associated protein, is important for protein synthesis in Saccharomyces cerevisiae under nutritional stress conditions. *J Mol Biol* 358(4):1023-1031.
235. Griac P & Henry SA (1999) The yeast inositol-sensitive upstream activating sequence, UASINO, responds to nitrogen availability. *Nucleic acids research* 27(9):2043-2050.
236. Schreve JL & Garrett JM (2004) Yeast Agp2p and Agp3p function as amino acid permeases in poor nutrient conditions. *Biochem Biophys Res Commun* 313(3):745-751.

237. Grousl T*, et al.* (2009) Robust heat shock induces eIF2alpha-phosphorylation-independent assembly of stress granules containing eIF3 and 40S ribosomal subunits in budding yeast, Saccharomyces cerevisiae. *J Cell Sci* 122(Pt 12):2078-2088.
238. Rogelj B*, et al.* (2012) Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Sci Rep* 2:603.
239. Ghaemmaghami S*, et al.* (2003) Global analysis of protein expression in yeast. *Nature* 425(6959):737-741.
240. Favre A, Moreno G, Salet C, & Vinzens F (1993) 4-Thiouridine incorporation into the RNA of monkey kidney cells (CV-1) triggers near-UV light long-term inhibition of DNA, RNA and protein synthesis. *Photochem Photobiol* 58(5):689-694.
241. Levin JZ*, et al.* (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 7(9):709-715.
242. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
243. Yassour M*, et al.* (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 106(9):3264-3269.
244. Nagalakshmi U*, et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320(5881):1344-1349.
245. Hofacker IL*, et al.* (1994) Fast Folding and Comparison of Rna Secondary Structures. *Monatsh Chem* 125(2):167-188.
246. Darty K, Denise A, & Ponty Y (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25(15):1974-1975.
247. Bailey TL & Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.
248. Alexa A, Rahnenfuhrer J, & Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22(13):1600-1607.
249. Shi Z, Montgomery TA, Qi Y, & Ruvkun G (2013) High-throughput sequencing reveals extraordinary fluidity of miRNA, piRNA, and siRNA pathways in nematodes. *Genome Res*.
250. Fischer SE*, et al.* (2011) The ERI-6/7 helicase acts at the first stage of an siRNA amplification pathway that targets recent gene duplications. *PLoS Genet* 7(11):e1002369.
251. Welker NC*, et al.* (2011) Dicer's helicase domain discriminates dsRNA termini to promote an altered reaction mode. *Mol Cell* 41(5):589-599.
252. Scherrer T, Mittal N, Janga SC, & Gerber AP (2010) A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PLoS One* 5(11):e15499.

253. Tsvetanova NG, Klass DM, Salzman J, & Brown PO (2010) Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae. *PLoS One* 5(9).